FAKULTÄT FÜR !NFORMATIK

Faculty of Informatics

Web Mining in Online Communities

A Comparison of Web Crawling and Mining Tools and their Economical Appliance

MASTER THESIS

Submitted in partial Fulfilment of the Requirements for the Degree

Master of Science

in

Business Informatics

by

Berndt Winter
Matriculation number 0209759

at the
Faculty of Informatics at the Vienna University of Technology

Supervision:
Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Assistance: Univ.-Ass. Mag. Julia Neidhardt

Vienna, April 2014

(author signature)

(supervisor signature)

FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

Web Mining in Online Communities

Eine Gegenüberstellung von Web Crawling und Mining Software und deren Anwendung im wirtschaftlichen Bereich.

DIPLOMARBEIT

Zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Business Informatics

eingereicht von

Berndt Winter
Matrikelnummer 0209759

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung:
Betreuer: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Mitwirkung: Univ.-Ass. Mag. Julia Neidhardt

Wien, April 2014

(Unterschrift Verfasser)

(Unterschrift Betreuer)

Technische Universität Wien
A-1040 Wien Karlsplatz 13 Tel. +43-1-58801-0 www.tuwien.ac.at

Erklärung zur Verfassung der Arbeit

Berndt Winter
Weingartenweg 6, 2751 Matzendorf

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum)

(Unterschrift Verfasser)

Abstract

Web communities are about one of the most comprehensive data sources of our time. Hundreds of millions of users are discussing people, political issues, product characteristics and many other topics in an unguided way making online communities a potentially valuable resource for market researchers. There is a large amount of tools available for extracting and analysing web data. The big number of available tools makes it difficult for any web researcher to choose the tool fitting his needs best. The vast amount of existing web mining tools with strongly differing capabilities leave the prospective user disoriented. The present thesis addresses this issue by creating a comparison and overview of such tools. The process to achieve the desired overview starts by discussing and defining the information which can be extracted and derived from online communities using web mining. The extracted data should be valuable for market research purposes. In a second step selection of tools is made based on the theoretical abilities each tool can offer. Tools, theoretically and practically suitable for web based market research, are chosen and used for web crawling and mining tests in a third step. The results will be screened and analysed to ensure their accuracy. Readers of this thesis will learn how web mining in online communities can be conveyed. They will find a comprehensive overview of mining tools and a comparison of their abilities. They will learn about those tools suited best for accomplishing web mining in online communities and see how freeware tools are performing in practical tests.

Contents

1. Introduction and Goal of this Thesis

Let us begin with an example: A company selling mobile phones wants to examine people's opinions about their products. They could use traditional analysis techniques such as opinion polls, conduct surveys via telephone or hand out questionnaires. Often surveys are carried out with customers after their purchase or with volunteers who receive a small token like the participation in a sweepstake. The disadvantage of these techniques is obvious: In such a supervised, artificially created environment the interviewees tend to be biased on their opinions and will certainly argue differently than they would do in an anonymous environment.[1] The World Wide Web is such an anonymous environment. Customers all over the world share their opinions about products, companies, trends, people in public and all sorts of things in blogs, forums, product ratings and others. The mobile phone distributor can take a look at websites like Amazon selling their mobile phones and offering a product rating platform open to anyone for an insight customer view on their phones. The phone company now has the option to let its staff read each review and list the pros and cons mentioned in those reviews. This is possible but very tedious, time consuming and expensive. There exists another, more automated way to extract the essence of those reviews: web mining in online communities.

In the almost infinite web space and in libraries all over the world exists a considerable number of books and research papers applying to web mining and some others tackling the issue of online communities. What is missing so far is an in-depth judgement on existing data mining tools in respect to their online mining abilities. This thesis shall elucidate mainly three problems. First it will focus on what kind of useful information existing data mining tools can gather from collecting and analysing data in online communities like blogs, forums and product rating platforms. The community types are compared on that basis and appropriate ones containing the desired information are selected. Secondly it will be examined which tools are suitable for web data extraction and analysis tasks by comparing their abilities. The tools appropriate for data mining are selected. The third most challenging task is to use those tools on examples for web data mining tasks, and examine how accurate the gathered results are. Results of selected tools used on chosen communities are compared with each other. Solving those problems should allow the reader to gain thorough insight in the topic of web mining, get to know what tools will fit the needs to conduct specialized mining tasks for all three commonly used web mining types: How they can be installed and used, how possible results look like, and how accurate they are compared to manually completed analyses. Most importantly, readers will get to know which tools work best for the proposed web mining tasks and whether free software is sufficient or commercial software is likely performing significantly better.

1.1. Research Problem – Questions to be answered

Normally one or two research questions are addressed within a master thesis. The essence of this thesis problem statement can be summarized in one question with a wide focus:

"What existing data mining tools are suitable to perform web mining in online communities for market research purposes and how do freely available tools compare in practical use?"

In order to split the large volume of work that is required to answer this comprehensive question the formulation of the question is split into six subquestions which are equally milestones for the work progress and the answers are given sequentially from chapter to chapter. Those subquestions are in detail as follows:

1. **What type of information interesting for market research purposes could you explore with web mining tools?** This question is answered by extracting and summarizing knowledge from related research work.
2. **Which communities are suitable for web mining data sources?** Based on the knowledge gathered answering question 1, a number of criterions defining an ideal community for data crawling and mining will be specified. Two communities are going to be selected that meet the criterions and have content stored in a way that is structured conveniently and easily extractable.
3. **Which information can you gather through web crawling and mining?** For each of the two selected communities three questions will be defined as well as their corresponding answers. Each question for a community addresses another data mining type - one for structure mining, one for classification based content mining and one for lexicon based content mining. The questions will be based on related research work in that area. The practical test phase in chapters 5 and 6 will use mining tools to answer those questions.
4. **Which available tools are suitable for crawling and mining data necessary to answer the questions defined in the first place?** The tools have to meet functional criteria. Suitable tools will be selected for practical test runs. Commercial software is included in an overview and compared to freeware tools but will not be used or discussed further.
5. **How do existing data mining tools prove themselves in real-world application? How do they compare to each other?** A set of freeware tools like open source software, trial and community versions of commercial products are going to be installed and tested on the preselected community sites to answer the six questions defined earlier. Basic tool functions will be presented as well as tool abilities and the collected results will be rated and classified.
6. **How precise are the results? Do the tools gather information on a site correctly or not?** The results collected while answering question 5 will be reviewed, benchmarked in terms of precision and recall[1] as well as other evaluation criterions by comparing them manually with the actual content on the website. The quality of results using different tools is compared.

1.2. Expected Results

By answering questions 1 and 3 mentioned above this thesis will give insight into the area of online market research to online communities. The reader will get a basic understanding of the sort of information gathered from certain communities and what type of market research problems one can address by conducting online data mining. Those results can be viewed in chapter 4.

The reader will learn in chapter 4 which type of online community types exist and which of them are suitable for market research purposes.

Besides the theoretical market research aspect the main focus of this thesis lies on feature analyses and application of data crawling and mining tools. The first result emerging from a comprehensive analysis of software feature data found on websites and in product manuals is an overview table. The table states which product has a desired feature and which one not. Each program is tested for its ability of answering any of the six questions as defined above, as well as its prize and where it can be found. This comparison answers question 4 and can be found in chapter 4.

The actual mining task and its results are documented for each tool used leading to comprehensible descriptions of how those tools basically work, how they are utilized, what results could be obtained in

[1] Those terms are explained in chapter 3

terms of answers to the questions and how correct and valuable those results are. The collection of tool documentations as well as a table summarizing the results can be found in chapters 5 to 7.

1.3. Methodology and Structure

Speaking of methodology, conducting market research in the internet generally requires the following steps to ensure a structured approach [1]:

- Specify marketing research questions and identify appropriate communities.
- Collect and analyse data

Other points of interest for market researchers not further discussed during this thesis are:

- Assurance of user trustworthiness and validity of user text provided
- Respecting research ethics by not doing harm to communities whilst publishing adverse results the community does not go along with.
- Checking the research results with the authors of the reviewed content.

Theoretical work starts with a comprehensive search for related work to get an overview of state of the art research papers. These papers will be presented in chapter 2 and build a base for the questions that will be defined later and are strongly tied to similar questions that have already been answered in earlier research work.

Based on literature research chapter 3 follows explaining the basics of data and web mining including the three types of web mining: structure, content and usage mining. Other important terms and concepts like sentiment analysis or opinion mining are explained and will be used later in the practical part. In addition the main characteristics of online communities as well as the different types of existing communities are described. This step is very important to get knowledge of web mining and community domains and further know-how carrying out the following steps.

According to the usual market research practical steps start by defining the relevant questions and finding appropriate communities which provide qualified content to answer the questions given. Since research work of this thesis does not seek information regarding a specific product or topic, it is possible to exchange the sequence of these tasks. Consequently chapter 4 starts by seeking communities well worth examining and containing valuable information which can be used for data mining tasks as seen in related work. Appropriate community criteria are defined for conducting the desired web mining task. Two communities are selected regarding these criteria. Afterwards questions for those communities are formulated that will be answered in the software test chapter by using the mining tools. For each community three questions are asked – one regarding structure mining and two regarding content mining – summing up to a total of six questions.

Besides community choice, chapter 4 is dedicated to selecting the right data mining and crawling tools. Firstly, comprehensive internet research is done seeking an extensive selection of available crawling and mining tools. Secondly, criteria are defined which a tool has to meet in order to qualify for the following practical tool evaluation step. Thirdly, the tools are compared to each other regarding their features. This comparison is summarized in a table. Those tools meeting the features necessary for conducting web mining are selected according to the next steps required.

After communities and tools have been selected in chapter 4, chapters 5 and 6 start with the actual crawling and mining tasks. As figure 1.1 demonstrates, the mining process can be regarded as a simple input-output process chain where the community websites containing HTML code are the raw material that is extracted and processed by each crawler. Each crawler produces a different result depending on its capabilities and settings used for data extraction. Those results are reviewed and compared to the actual website content. The most precise, complete and clean results form the new raw material that feeds the mining tools. Those mining tools are used to process, transform and visualize the data in a way to answer previously defined questions. It is possible that one single tool supports both crawling and mining.

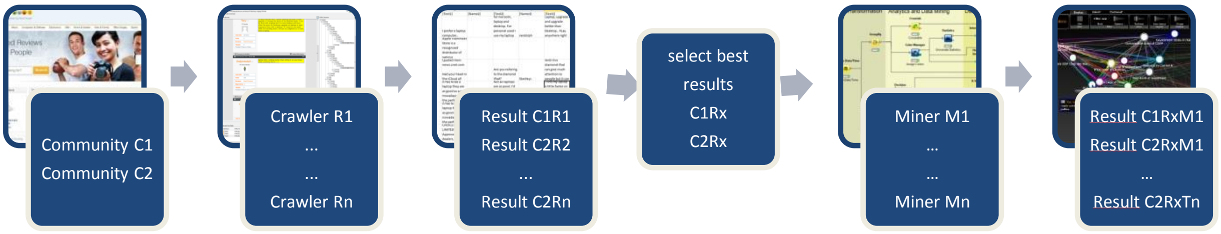


Figure 1.1: Process flow diagram of this thesis' practical research work. Website content of two communities will be crawled and mined with tool after tool. This leads to various results that potentially allow answering the six predefined research questions.

Each tool is reviewed in detail, the function principle is explained and crawling or mining results are presented. Each result is compared to manually extracted and analysed results, evaluating precision, recall and overall quality of the result. The results are summarized in an overview table that shows which tools support crawling, structure or content mining and how the results compare to each other.

At the end of chapters 5 and 6 the results gathered from crawling and mining tasks are summarized and compared in structured tables.

Chapter 7 finally closes this thesis with a summary of all findings by answering the research questions using new insights gathered during the research work done in the course of this thesis.

2. Related Work

Since the main literature work on this thesis was done in the end of 2012, most references listed here are from 2012 or older. More recent work may have appeared in the meantime but will not be covered in this chapter. Some terms like lexicon or machine based mining mentioned here will be explained in further detail in chapter 3.

Research so far has been done on various data mining software by testing and comparing their technical abilities in [2]. Unfortunately this analysis is a bit outdated - it was released 16 years before this thesis is written. Another paper describes mining software that is specialised on analysing web logs [3]. This software is useful for web usage mining that will not be focused on during this thesis. In [4], the author provides a general overview of web crawlers and classifies them based on the extraction technique they use. [5] compares user interfaces and technical capabilities of 19 Data Mining tools to each other. As the other comparisons it is rather old too (15 years).

The most recent related work mentioned here is [6]. It compares crawling functionality including some tools not mentioned in this thesis. These tools are Screen-scaper, Automation Anywhere 6.1, Web Info Extractor and Mozenda. Most of them offer free software trial versions. The comparison is done only rudimentary mainly stating if a tool can process structured or unstructured data. Since this paper was released in June 2013 after the tool test phase was finished, tools mentioned there were not considered for this thesis.

A bit more up-to-date is [7], which describes an attempt to do structure mining on Japanese communities discussing gender related topics using the tool “Companion“ that is not freely available on the web and therefore could not be included in this thesis. Community analysis is both done synchronic for static view on community network connections and diachronic to visualize community development over time using historical data from the University of Tokyo.

[8] is a recently published paper that compares different algorithms for community finding on Facebook with the main purpose to tests a newly developed community finding (structure mining) tool named iLCD. That tool and the underlying community finding algorithm were presented earlier in [9]. The algorithm works with a multi-agent system[2] that replays the evolution of a network. It should overcome the weaknesses of traditional algorithms like the impossibility to deal with dynamic networks, noisy results, failing detection of overlapping communities and exponentially increasing runtimes on larger scale projects. iLCD is compared to other tools as well as the native list of friends provided by Facebook and is able to outperform them in the shown test scenarios.

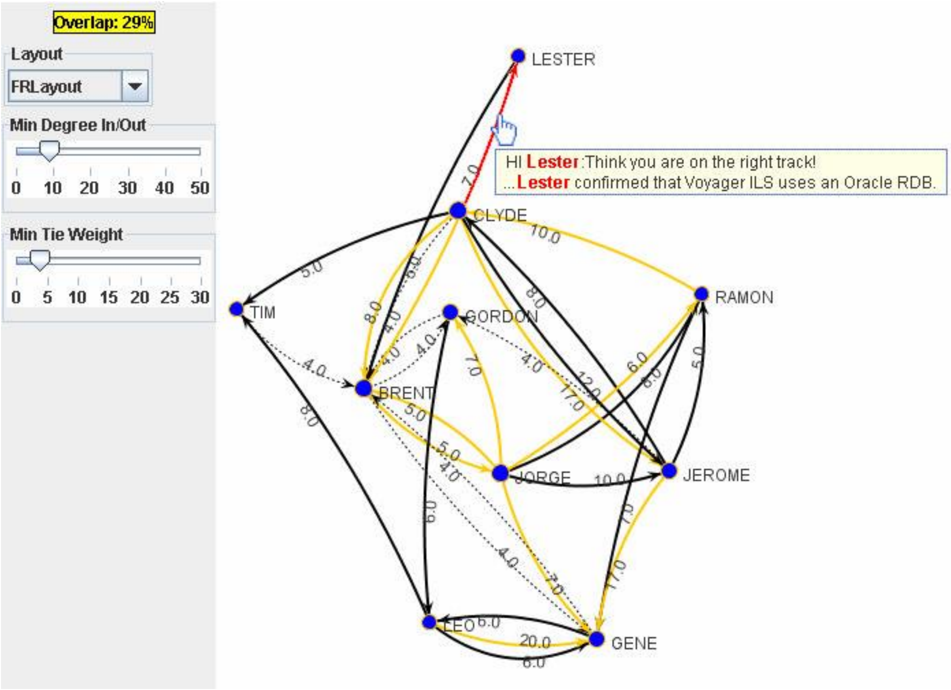
Another literature covering this topic is [10]. The author (Mikolaj Morzay) was very active in searching and developing his own data mining techniques, algorithms and tools since 2002, especially those appropriate to mine data in online communities. This book, which in fact is a habilitation thesis, is mainly dedicated to data mining in three different kinds of social platforms – blogs, internet forums and online auctions. Unfortunately the whole book seems to be nowhere available in Austria at the time this thesis was written.

However some research papers from the same author could be obtained. These papers formed the base of his knowledge put into the thesis. [11] and [12] describe a new data mining technique that searches for suspicious patterns in sellers’ ratings on the online auction Allegro. An algorithm is introduced that

[2] A system containing of several small, to a certain degree autonomous acting computer programs (= software-agents) which collectively solve a problem.

counts transactions as links between buyers and sellers and excludes auctions below a certain sales price. Cluster recognition allows reduction of fraudulent ratings and makes the rating system overall more trustworthy. [13] too addresses the feedback-trustworthiness in online auctions. The proposed method takes into account the number of buyers that did not give a feedback because a non-feedback may occur in fear of a negative vengeance feedback making it relevant for member rating. [14] is dedicated to an algorithm that detects fake feedbacks from fake users and lowers the trust-ratings of users giving or receiving such ratings. A fake user can be someone that gives ratings exclusively to one single seller and wins auctions with overall very low sale prices. Also cliques of users that invariably give ratings to each other are considered as fake. In [15] once again a method for evaluating trust and distrust on sellers in online auctions is presented, which this time is (seller) graph based. It especially gives higher weight to buyers that mostly bought products from sellers that are not linked and have no commonalities to the seller under review which leads to a better experience base for seller-quality-comparison.

[16] describes a tool under development named Internet Community Text Analyzer (ICTA). This tool is able to search online discussions (e.g. in forums) for links between discussing participants. This is done by counting and displaying messages that contain other user names. This evaluation, which is presented as a labelled graph, should enable faculty and administration to get deeper insight into students' and teachers' usage of e-learning platforms and their communication among each other with the intention to enhance the e-learning experience in future classes.



Overlap: 29%

Layout: FRLayout

Hi **Lester**:Think you are on the right track!
...**Lester** confirmed that Voyager ILS uses an Oracle RDB.

Figure 2.1: ICTA, described in [16], returns a directed graph that shows interaction intensity between users in an online discussion community. Vertices represent users while edges and their labels represent the amount of information exchanged. The sliders on the left allow excluding weak ties alias links with only few information exchange. By moving the mouse cursor over a link the message sent is displayed.

The authors of [17] describe an early state of an opinion mining tool that is evaluated on twitter statements. The goal of this development is to make the tool recognize opinions that fit to a specific topic, evaluate them as positive or negative and show their development over time. It should also recognize key personalities that represent these opinions and how their demographic user data are. The paper too describes difficulties that have to be considered while programming opinion mining software.

In [18] the Flickr API is used to crawl and analyse social networks and photo distribution on Flickr. One main discovery they made with their mining work is that popularity of photos is ascending steadily meaning that the number of people newly discovering a photo in a given period of time is remaining almost constant – except for events as the propagation of a picture on the Flickr start page which causes a sudden increase. The authors also discover that most people liking a specific picture are either direct friends or friends of friends of the owner meaning that popularity is limited to a relatively small social group.

In [19] an application is developed with means of the Facebook API that is able to visualize the development of interests and friend lists of certain persons over a period of time. The way friendships are influencing interests can be observed too. Each Facebook-member under review has to accept the terms of such an application in order to allow it reading personal data which makes this approach difficult to use for larger user groups.

[20] describes a lexicon-based approach for extracting sentiment from a text. The proposed Semantic Orientation CALculator (SO-CAL) determines sentiment taking into account valence shifters like negotiations or intensifiers. The authors show that the proposed, lexicon-based method is robust across multiple domains by testing it with various reviews on the product rating platform Epinions with different product reviews. The SO-CAL, unlike some other lexicon-based methods, takes into account not only adjectives but also verbs, nouns and adverbs. Every term has a value from -5 to +5 making fine graduations of strong and weak sentiment terms possible. Intensifiers can raise (e.g. very) or lower (e.g. somewhat) those values.

[21] is a lexicon based document sentiment classification approach that does not use a precompiled sentiment word lexicon. Instead the proposed AMOD approach utilizes a search machine to automatically extract a training set for blogs focussing on a specific domain. Test runs in this work focus on the domain “movie” to classify each 1000 positive and 1000 negative reviews. By searching for adjectives in the extracted training set a domain specific sentiment lexicon is created. Lexicon quality is further improved by considering polarity-changing words as “not” or “neither” as well as seed word proximity meaning only those adjectives were taken into account that are placed close to the feature term (in this case “movie”). Tests shows classification results with such a lexicon to be superior to a supervised machine learning method.

[22] proposes a lexicon based text mining approach to overall classify review sentiments. It counts the number of positive and negative terms in a document. If the number of positive terms is greater the review is considered overall positive and vice versa. The method makes use of valence shifters.

[23] presents techniques to mine and summarize product reviews and present the features mentioned as well as their sentiment direction. Product features are determined using association mining to find commonly used feature terms in reviews. For sentences containing frequent features the nearest adjectives are considered as their effective opinion.

[24] introduces a web service for extracting and summarizing features and their sentiment orientation on a hotel review site. The proposed tool uses ontologies to identify key features. Similar to [23] the tool counts specific adjectives occurring directly before or after the feature term and as measurements for the sentiment direction.

[25], [26] and [27] all discuss sentiment classification with machine learning methods. [25] compares result accuracy between the three different classifier algorithms Naive Bayes, Maximal Entropy and a

Support Vector Machine[3] as well as between using feature unigrams (consisting of 1 word each, e.g. fun), bigrams (2 words, e.g. not fun) and part of speech (POS) tags (e.g. they say it is not fun but I disagree). Results show that differentiating just the presence or non-presence of a unigram leads to significantly better results than counting the number of occurrences. Overall best results could be obtained using unigrams with Support Vector Machines. [26] includes contextual information by analysing complete phrases instead of single terms to enhance classifying results. This contextual information includes negotiations (e.g. not good), intensifiers (e.g. very good), negotiation phrases that work as intensifiers (e.g. not good but amazing) and more. The process is done in two steps. First polar expressions are separated from neutral one and then the polarity of selected expressions is determined. [27] builds upon [25] and uses the same classifying algorithms and feature variations. It presents a Twitter tweet search and sentiment classification application that can be executed and tested under `http://www.sentiment140.com/`. It uses a training set labelled by emoticons used within the tweet messages. ☺ labels a message as positive and ☹ as negative.

2.1. Relation to this Thesis

Although the software comparisons mentioned in this chapter are rather outdated they are still helpful to form a basement for the tool comparison table in chapter 4.3 since most tools mentioned there still exist although user interfaces and functionalities have changed since then.

ICTA was used as guideline for formulating the structure crawl question for the product forum as presented in chapter 4.2. The result ICTA produces is of a very similar type as the result obtained by answering this question.

Mikolaj Morzays work about rating authenticity evaluations had considerable influence on the structure crawl question for Epinions presented in chapter 4.2 as well. Morzays work describing the process of mining the structure of ratings in online auctions by analysing who gave ratings to whom can be used to determine the trustworthiness of ratings. The task defined for Epinions analyses the rating structure in a very similar way.

A lot of work about lexicon and machine learning based sentiment classification methods are used as basis to get an idea how to formulate content mining questions and how to conduct content mining tasks. Lexicon-based sentiment analysis is done in a more simplified way compared to [20] or [23] since this thesis focuses more on the data mining process itself and less on perfectly accurate results. Therefore dictionary creation and valence shifters are not considered. The machine learning method papers show similar classification tasks as this thesis uses for machine learning based content mining. Again this thesis uses simplified methods of those mentioned in the research papers.

Work on Twitter, Facebook and Flickr discusses structure mining similar to the task performed during this thesis but extend it to an additional time dimension. This means they are showing the development of links over time. This dimension is not covered by this thesis' practical part and therefore an interesting starting point for further reading on the structure mining topic.

Further influence on this thesis had [21] which uses precision and recall (or, to be exact, a score computed with a formula including them) as measurements for classification correctness and made use of the simple lexicon based classification method by counting positive sentiment words and subtracting negative ones to get a positive or negative result in a similar way as it is done in this work.

[3] Naive Bayes and SVM are described in chapter 3.4.3., page 17

3. Basics - Algorithms and Methods

3.1. Data Mining

Since web mining is a subdomain of data mining it makes sense to begin with an introduction to data mining because the basic concepts of data mining apply to web mining as well. Data mining is defined as the process of discovering patterns or knowledge in data sources like databases, plain text, pictures, the web etc. The discovered patterns have to be valid, usable and comprehensible. [28]

The most common data mining methods are [28]:

- **Supervised learning** probably is the most commonly used method for data or web data mining. It is also known as classification and is used to divide information in different classes or categories. By manually classifying a set of training data a classification function can be learned that should be able to automatically classify similar data sets.
- **Unsupervised learning** does not use manually classified data. Instead a learning algorithm tries to find hidden structures or laws in the data set. An unsupervised learning method that is predominantly used for web data mining is clustering. It forms a given amount of clusters with similar data in each cluster. Clustering can, for instance, classify websites into groups where each group represents a certain topic.
- **Association Mining** searches for data combinations that occur regularly together. Web mining can use association mining, for instance, to find rules regarding the behaviour of users visiting and buying stuff from a site.
- **Sequential pattern analysis** seeks data combinations that co-occur in a certain order. This analysis can help finding rules for the way users navigate through the website.

A data mining application usually starts by programming a data mining algorithm for a certain domain that is able to automatically find suitable data sources in the available bulk of data. Based on these sources data mining is conducted in three steps that are iteratively passed through one or several times [28]:

- **Preprocessing:** Raw data normally is not suitable for data mining. Data often contains unwanted parts like HTML tags, pictures, spam, etc.. Furthermore data sets can be too large and contain unneeded attributes, making data reduction per sampling and attribute selection obligatory. This step usually makes up 80-90% of total effort put in the data mining process. [30]
- **Data Mining:** The preprocessed data is forwarded to a data mining algorithm that extracts patterns and knowledge from it.
- **Postprocessing:** A lot of data mining applications discover patterns that are not utilizable. During this step different evaluation and visualization techniques are used to identify the desired patterns.

3.2. Mining the Web

In contrast to conventional Data Mining which is conducted on structured data stored in relational databases and tables Web Data Mining uses the world's largest, freely accessible database that is way more heterogeneous and less structured – the web. This source has some attributes making information collection a difficult but fascinating task as well. Some of them are [28]:

- The amount of data is vast and steadily growing. Information about more or less every topic in existence can be found on the web.
- The web features a multitude of different data representation types. Such types include structured tables, semi-structured websites, unstructured texts and multimedia files (pictures, audio and video)
- Information on the web is heterogeneous meaning the same information can occur on several websites in different representation forms or modified formulation. This makes consolidation of information stored on different sites a challenging task.
- A large amount of information on the web is linked to each other. Hyperlinks can link to a sub-site of a website or to external sites. Sites that are linked to from several external sites can be considered as important, high quality sites. Modern search engines usually take this into account. E.g. the PageRank algorithm that forms the basement of Googles search engine counts every link from a site x to another site y as a vote of x for y. More votes equal higher relevance of this site. PageRank is described in further detail in chapter 3.4.1.
- Information from the web is "noisy". A typical website does not only contain the desired main content but also navigation links, banner ads, pictures, copyright- and privacy statements, etc.. This "noise" has to be removed in order to conduct a specific analysis. Additional "noise" is created due to the fact that information can be added to the web by anyone owning a computer with internet access nowadays and content often does not underlie quality controls resulting in bad quality and sometimes even completely wrong statements.
- The web offers services. Most commercial sites allow users to execute useful operations like buying goods online, fill out forms or pay bills.
- The web is dynamic meaning information changes constantly.
- The web is a virtual community. It is not only about data, information and services but also about interaction between people, organisations and automated systems. One can communicate with people all over the world practically without any delay and announce ones opinion about practically any topic in forums, blogs and review sites, also known as online communities.

Web mining works very similar to data mining. The main difference occurs during the data collection process. While traditional data mining usually starts with already collected data stored in a data warehouse[4], data collection during web mining can make up the lion's share of time and effort. This is especially true for Web Mining techniques necessitating crawling through a large bunch of websites. Web crawlers are discussed in the next chapter.

3.3. Web Crawling

Web crawlers, also known as spiders or robots, are programs used for automatic downloads of websites. Since the web is not a static resource but exposed to constant changes, crawlers are of high importance for applications helping them to automatically maintain information up to date by tracking changes.

Web crawlers can be used for various applications. For instance, an organisation can collect data from the webpage of a business rival. Another domain is supervision of websites, notifying users whenever a new piece of information occurs in the area under review. Besides, crawlers can be used for some

[4] A data warehouse is a database consisting of data originating from different sources that are represented in a consistent format.

dubious kinds of applications like collecting E-Mail addresses for the purpose of sending spam, collecting personal data for phishing or other forms of identity theft.

Commonly used applications of crawlers are search machines. They are used to build up an index for their search engines. This index is important to deliver answers to queries efficiently and quickly. The drawback of using crawlers for index build-up is the high demand of internet bandwidth. This high demand can result in server lockups due to too many page recalls. The crawler would more or less conduct a denial of service (DOS) attack making the site inaccessible for real human users. To get rid of this issue crawlers have to follow the crawler etiquette - a number of behaviour rules limiting how often a site can be accessed per time unit, instructing crawlers to identify themselves in the user agent HTTP-header with their name and version number and lastly asking crawlers to read the robots.txt file on the webserver that contains a list of sites forbidden to access for crawlers.

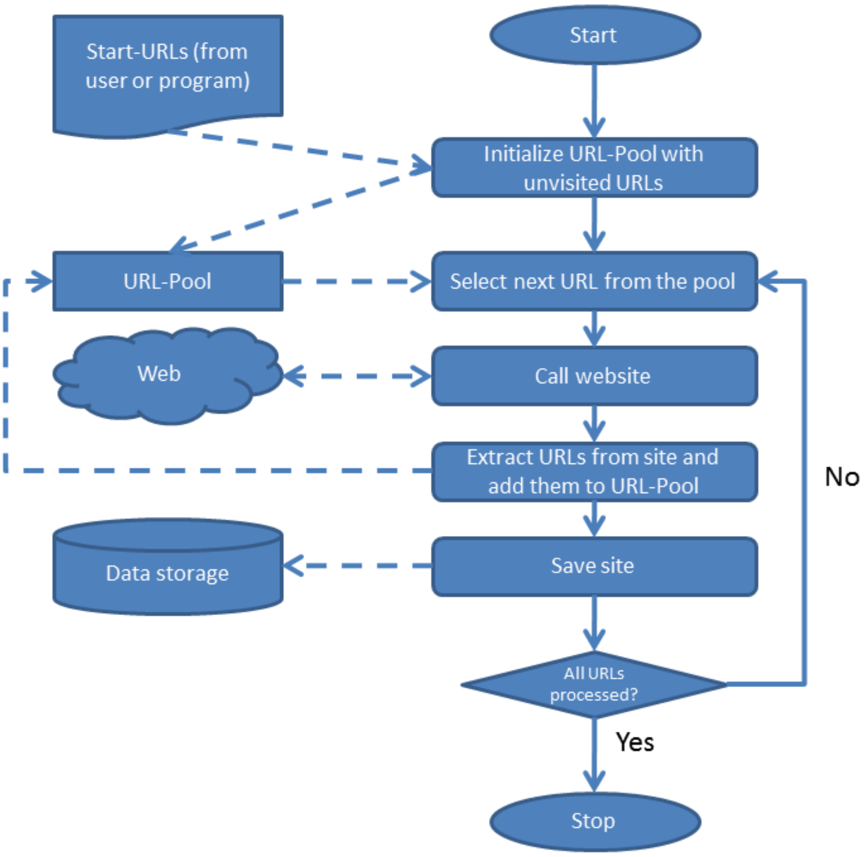

Start-URLs (from user or program)
Start
Initialize URL-Pool with unvisited URLs
URL-Pool
Select next URL from the pool
Web
Call website
Extract URLs from site and add them to URL-Pool
No
Data storage
Save site
All URLs processed?
Yes
Stop

Figure 3.1: Function principle of a very simple crawler form [28]

A very simple, sequential crawler form is shown in figure 3.1. Such a crawler uses resources rather inefficiently since just one of the three available resource types network bandwidth, CPU and storage transfer rate is used at a time. More complex crawlers exist that are able to simultaneous charge more available resources to capacity.

A lot of research work is done on crawler domain searching for increasing resource charge and higher quality result sets. A common method to measure quality of a crawler is the use of the two classification numbers precision and recall that are calculated as shown in the algorithm below [29].

$$\boldsymbol{Precision = \frac{number\ of\ relevant\ sites\ a\ crawler\ finds}{number\ of\ all\ relevant\ sites\ on\ the\ web}}$$

$$\boldsymbol{Recall = \frac{number\ of\ relevant\ sites\ a\ crawler\ finds}{number\ of\ all\ sites\ a\ crawler\ finds}}$$

Algorithm 3.1: Crawler quality evaluation metrics precision and recall

The goal of any retrieval algorithm should be to maximize both values. Normally algorithms providing higher recall lead to lower precision since getting larger result sets increase the chance for them to contain false positives.

3.4. Web Data Mining Types

As soon as data has been collected, the same three-part process used with data mining is conducted including preprocessing, data mining and postprocessing. However, the techniques used with every step can differ quite clearly from traditional data mining.

Based on data types used for the mining process, Web Mining applications can be divided in three categories. [28] Text mining as an important subtype of (web and other forms of) content mining will be explained in this chapter as well:

- Web structure mining
- Web content mining
    - Text mining
- Web usage mining

3.4.1. Web Structure Mining

Structure mining explores the hyperlink-structure on the web. It is helpful, if you want to know how web sites are related to each other or in which way users are in contact with each other.

The two most famous search algorithms based on a structure mining approach are HITS (Hyperlink-Induced Topic Search) and PageRank. Research on structure mining based approaches began in 1996 when it became obvious that content based methods just looking for textual similarities were not sufficient to deliver appropriate search results to the user. For instance, searching Google for “Web Data Mining” in 2009 revealed over 14 million results. Since people usually only look at the first 30 to 100 results ranking of search results became inevitable.

Adapted and improved versions of HITS and PageRank are still used in popular search engines like Google, Yahoo and MSN. HITS was introduced in January 1998 by John Kleinberg and works with hubs and authorities. Hubs are sites that contain many links to high quality sites and authorities are websites that many hubs link to. HITS ranks sites according to their hub and authority score enabling search machines to display more relevant, higher quality sites first. [28]

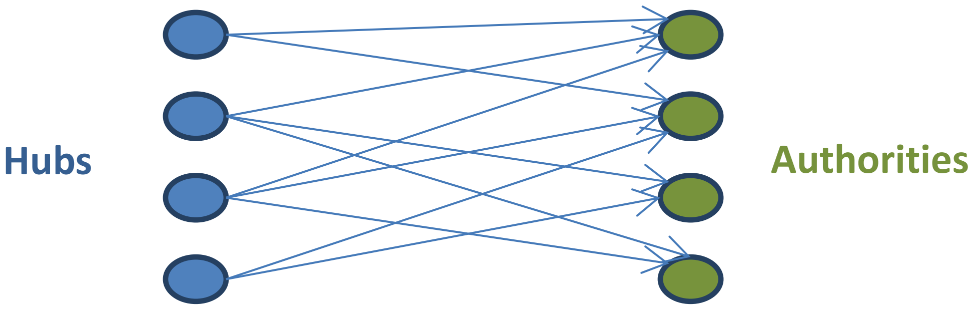


Figure 3.2: Function principle of HITS: Hubs (HQ link collections) link to authorities (HQ sites)

PageRank was introduced in April 1998 by Sergey Brin and Larry Page. Based on this algorithm they designed the search machine Google. PageRank relies on the democratic nature of the web by regarding the web link structure as site quality indicator. This is done by counting every link from site A to site B as vote for A to B. In addition to that votes are weighted in a way that votes from important, high quality sites count more than those from unimportant having only few votes for themselves. [28]

Besides ranking search results Hyperlinks can also be used to find web communities. A web community is a net consisting of sites densely linked to each other indicating that they are hosted or visited by people with similar interests.

Regarding websites as actors and their links to each other as relationships between them allows finding virtual web communities that can be displayed as networks or graphs. This procedure is called social network analysis.

Communities are not limited to exist on the web but show up in emails and text documents as well. There exit algorithms for finding communities in all of that sources.

3.4.2. Web Content Mining

Web content mining addresses the data that can be found on a web page. Data on web pages is mainly textual, can be structured (e.g. a product price comparison overview on `www.geizhals.at`) or semi-structured (e.g. text on a newspaper site is structured by topics but the articles themselves mainly consist of unstructured text). It has to be filtered and sorted to gather some specific information from it. For instance, counting frequently occurring words in a text document while removing some stop words that lack entropy[5] (e.g. *the*, *and*, etc.) can give a rough overview of document content since describing a certain topic usually topic-related terms are used very frequently as shown in [30].

Structured data usually originate from underlying databases and are displayed in fixed patterns. Extracting and merging that data from various sources enables offering useful services. Those services include adaptable collection of information from the web, product price comparisons or meta searches that merge search requests from several search engines.

Semi-structured data largely containing of plain text can be found almost everywhere on the web. Other than structured data plain merging of data from different sources and storing the unstructured text fragments in data cells would not add much value here. The text fragments mostly are just too

[5] Measure for information content

long and contain too much diverse information for one single data point making further splitting and processing of these fragments obligatory.

Whatever data type you are collecting, the extraction process from websites always remains a challenging task. The extraction process is completed by storing gathered data in some type of structured data base or table making it further processable by data mining tools. Data extraction is done by crawlers as mentioned in chapter 3.3. but actual methods and algorithms the crawler uses can differ quite a bit. Structured or semi-structured data can be extracted in three different ways [28]:

- **Manual approach**: A human programmer searches the website and its source code for patterns. Then he writes a program able to extract demanded information. Because of its high effort this approach is only suitable for smaller amounts of websites.
- **Wrapper induction** is a supervised, semi-automatic approach. From a collection of manually labelled sites a program learns extraction rules enabling it to gain requested data from similar formatted sites.
- **Automatic extraction** is an unsupervised learning approach. From one or a set of given sites a program automatically finds patterns and grammatical constructs enabling the program to execute data extraction on similar sites. Since this approach works without any manual labelling it is applicable on huge amounts of websites.

Common claim among data miners is that 80 to 90 % of project work time is put in data preparation steps. [30] This holds true for text mining. What sets text mining apart from normal data mining is the type of processed data. While common data mining processes use mainly numerical data, text mining uses pure textual raw data. The main challenge therefore is to transform textual data in some numerical values that are statistically evaluable. To convert text in numbers, you count the number of occurring words or phrases in different ways like absolute or relative word frequencies or calculate a score based on the relevance of a term.

Before data conversion from text to numeric values can be performed, all or some of the following preprocessing steps are done [30]:

1.) **Choose the scope** of the text to be processed (documents, paragraphs, sentences, …)
While for clustering or classification entire documents are the proper scope, for sentiment analysis usually smaller text units such as paragraphs or sentences are preferable.
2.) **Tokenize**: Split text into discrete words called tokens.
For English text tokens can usually be separated through whitespaces or punctuation. Sometimes this does not work properly. E.g. U.N. or San Francisco would be separated in two tokens each splitting up the initial sense of the words. There are ways to overcome this problem. E.g. smart tokenization tries to detect abbreviations and avoids tokenization of these words.
3.) **Remove stopwords** (common words without entropy like *the, and*, etc.)
Removing stopwords leads to less storage demand and faster processing times while enhancing result accuracy since false similarities among these common words are ignored. Care has to be taken while analysing not only single words but whole phrases since removing stopwords can split up these phrases making them lose their meaning.
4.) **Stem**: Remove prefixes and suffixes to normalize words (e.g. *running*, *run* and *runs* are all stemmed to *run*)
Besides removing pre- and suffixes from words a dictionary is used for handling irregular word forms. For instance, *quickly* may be stemmed by removing the *-ly* suffix leading to the correct stem *quick* but *reply* must not be stemmed in that way since the result *rep* would not

make any sense. One very famous and frequently used stemmer is called Porter Stemmer introduced in 1980 that was later refined and extended to other languages by introducing the Snowball Stemmer. Stemming usually improves text mining results through reducing the word count in documents by grouping similar terms. Similarities between documents can much more likely be detected this way. One deficiency of stemmers is that they are unable to recognize the meaning of words leading to possible wrong processing of homographs[6]. For instance, *meet* can be used as a verb or a noun (i will *meet* my friend next Friday vs. everyone from the managing board will be at the *meeting*). A common stemmer will always stem the word to *meet* eliminating the distinction between them. To overcome this problem the advanced technique lemmatization was developed taking into account the context surrounding the word as well as the grammatical part of speech constructs leading to a more sophisticated „understanding" of the meaning of a word.

5.) **Normalize spellings**: Unify misspellings and spelling variations in one single token.
Such spelling normalization tasks are recommended when processing content of mediocre or unsteady quality like text taken from the web. Corrections can be done with dictionaries (e.g. equalizing british and US spelling of colour/color) or fuzzy matching algorithms that cluster together misspelled words.

6.) **Detect sentence boundaries**: Mark the end of sentences.
It is almost sufficient to count punctuation as sentence discriminators but as with tokenization abbreviations can contain punctuation and would split sentences making it necessary to use a few simple heuristics and classification techniques to accurately identify sentence boundaries.

7.) **Normalize case**: Convert all words to lower (or upper) case.
As stemming and spelling normalization this further reduces the word variety making analysis faster and more accurate.

3.4.3. Text Mining and Sentiment Analysis

Processing unstructured text fragments collected from semi-structured websites is done with several forms of text mining. This umbrella term involves many research fields as shown in figure 3.3 and can be used for texts originating from many more sources than the web including document and (research) paper archives.

[6] Words spelled the same way having different meanings. E.g. *lead* (metal) and *to lead* (to guide)

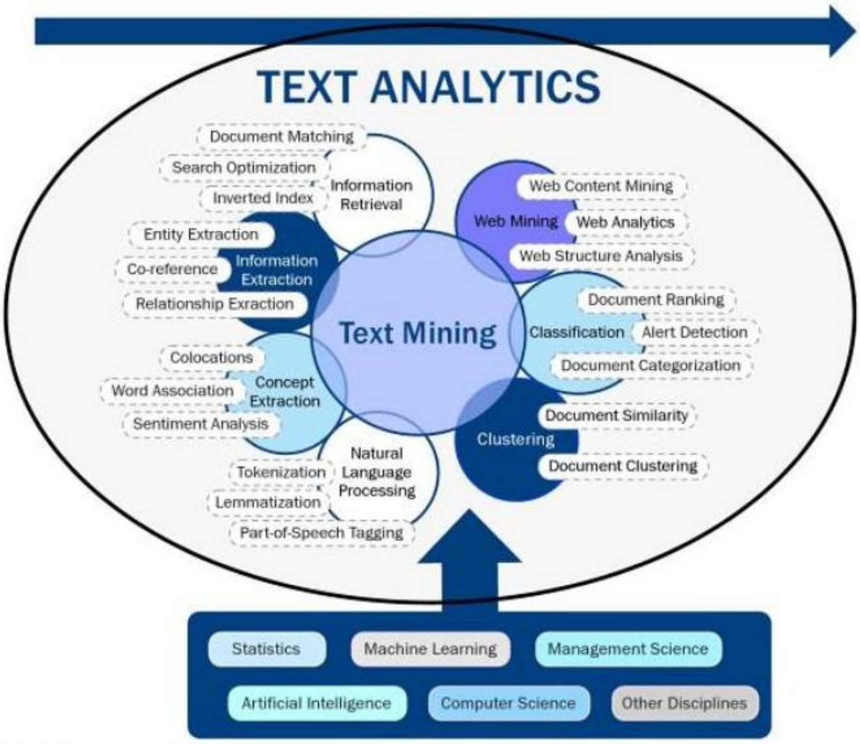


Figure 3.3: Text mining is based on knowledge from many external disciplines (blue rectangle at the bottom) and draws upon many text analytical components. [30]

For the scope of this thesis a special focus on the concept extraction technique sentiment analysis will be placed since this technique allows for classifying texts and sentences as positive or negative. Note that sentiment analysis is not strictly separated from all other text analysis fields but is using natural language processing techniques and information extraction techniques as well.

A definition stated in [20] says: "Sentiment analysis refers to the general method to extract subjectivity and polarity from text (and potentially also speech)." In other words sentiment analysis tries to determine if a text has a positive or negative meaning. Sentiment orientation on the other hand refers to the polarity and strength of words, phrases and texts determining which terms and phrases are considered positive and which negative. [31]

The following two methods for sentiment analysis exist [28]:

a) The **lexicon-based approach** uses dictionaries storing the semantic orientation of words and phrases.
b) The **text classification approach** is a supervised machine learning approach where a classifier is learned with labelled example texts.

Lexicon-based approaches (a) count how often words from the lexicon occur in the reviewed document. The resulting numeric values are further used to calculate the document sentiment. Dictionaries can be common or for a specific domain including typical sentiment words for a topic. E.g. for an iPod „good sound quality" or „stylish" will be typical positive phrases or terms.

Lexicon-based approaches depend on a dictionary that can be either created manually or automatically. The automated method uses seed words and automatically expands the list of words.

The lexicon itself usually consists mainly of adjectives with corresponding semantic orientation values that are usually either marked "positive" or "negative" or assigned values on a numeric scale. By counting occurrences of those adjectives in the document under review and summarizing their dictionary values the overall semantic orientation can be computed. For instance, a document that contains 55 adjectives marked as “negative” and 180 adjectives marked as “positive” is most likely a document with positive sentiment.

Usually term occurrences in texts are measured using inverse term frequency (TF-IDF). Instead of just counting any occurrence of a word in a single document the IDF value takes into account the number of term occurrences across the whole document set giving lesser weight on frequently occurring terms like “and” or “the”. The value for term *t* is computed as follows:

$$idf(t) = log(1 + (f(D) / f(d, t))$$

where *f(D)* is the number of all documents and *f(d, t)* is the number of documents containing term *t*.

Algorithm 3.2: Inverse term frequency (TF-IDF)

To enhance classification correctness of such lexica the effect of linguistic context has to be taken into account. E.g. the adjective “good” may not always have to be used in a positive way. It may mean the opposite by adding the negation “not good” or even be intensified by adding “very good”.

At least four types of adjective modifications can be distinguished [22]:

- No modification (e.g. “good”)
- Negated (e.g. “not good”)
- Intensified (e.g. “very good”)
- Diminished (e.g. “rather good”)

The **text classification approach (b)** can provide more sophisticated results on the trained topic but usually fails in gathering correct sentiment data from documents that are covering other areas. For instance, a classifier trained with movie reviews in [22] had a success rate of 85,1% with classifying other movie reviews but will most likely fail in correctly covering other topics such as reviews of hotel rooms because some of the learned phrases do not have the same positive or negative meaning in another context.

A frequently used, simple yet rather accurate and well approved algorithm for text classification is the (Naive) Bayes classifier. It classifies data sets into given classes by calculating the probability for each feature belonging to a certain class without taking into account dependencies between features. Just a small set 0 of training data is needed for adequate results. The calculation is shown in algorithm 3.3.

$$\boldsymbol{p(c_j|d)} = \frac{\boldsymbol{p(d|c_j)p(c_j)}}{\boldsymbol{p(d)}}$$

$$\sim \boldsymbol{probability\ of\ given\ instance\ d\ being\ in\ class\ c_j} = \frac{\boldsymbol{probability\ of\ given\ class\ c_j\ containing\ instance\ d * pobability\ of\ class\ c_j\ occurring}}{\boldsymbol{probability\ of\ instance\ d\ occurring}}$$

Algorithm 3.3: Naive Bayes classification formula for each class cj[7]

[7] Source: `http://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect_examples.pdf` (Mar 2014)

Another classification method is the Support Vector Machine (SVM). A SVM is a so called large margin classifier that classifies objects in a way maximizing the distance between class borders meaning the nearest objects belonging to two distinct classes should be as far apart as possible. Each object is represented as a vector. The simplest form is a linear SVM where a straight line can be drawn through the whole vector space to separate two classes.

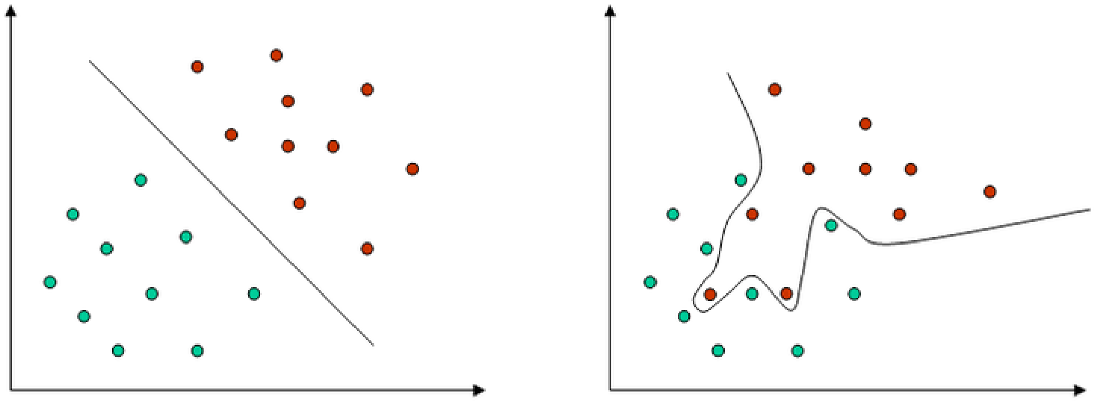


Figure 3.4: Linear (left) and non linear (right) support vector machine separation[8]

An important **application** of sentiment analysis in context of web mining is **opinion mining** where text content from reviews, forum or blog entries or other sources of belief from the web are collected. This task is technically difficult to implement but the results bear high value. E.g. companies are always interested in knowing what their customers think about their products and services. On the other hand potential customers like to know what product owners think about it before they make a decision for or against purchasing it. In addition opinion mining can deliver information regarding optimal placement of advertisements on websites. A product commercial will have better effect on a site containing many positive opinions about that product and may have virtually no effect at all if opinions on the site are mainly negative. In that case a commercial for a competing product may lead to higher success.

In general there are three ways in which opinion mining can be conducted [28]:

- **Sentiment classification** regards opinion mining as a text classification problem. The whole text under review can either be classified as positive or as negative.
- **Feature based opinion mining and aggregation** searches single sentences for details, mainly for which aspects of an object people like and which not. An object can be a product, a service, a topic or an organisation. In a product review this technique can identify features a reviewer writes about and classify them as positive or negative. For instance, in the sentence “battery runtime is too short” “battery” is the feature and it is classified as negative.
- **Comparative sentence and relationship mining** identifies sentences that compare one object to another. An example could be “The battery runtime of camera A is significantly shorter than camera B.”.

A problem influencing all those mining techniques is opinion spam, where dishonest opinions are published in order to make one’s own product look better than it is or cast a slur at the competitor’s

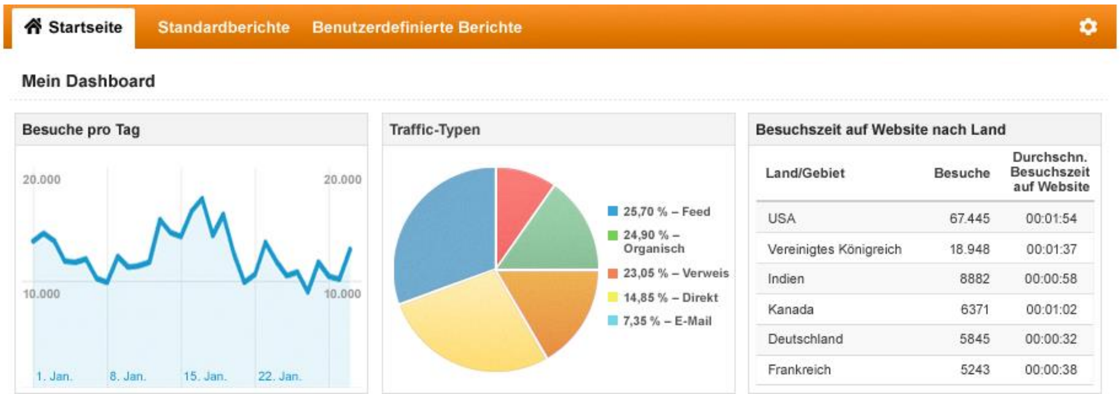
[8] Source: `http://upload.wikimedia.org/wikipedia/de/a/a0/Diskriminanzfunktion.png` (Mar 2014)

products. Due to the anonymity of the web reviewers' identities often cannot be determined making detection of spam a challenging task.

Besides sentiment analysis text content mining can be used commercially for warranty claim analysis (e.g. to automatically find common problems with a product that should be resolved as quick as possible), quintessence extraction from customer reports and for broader analysis of advertisements. Further customer complaints or warranty claims can be analysed to detect common and shifting opinions and values among them. Text mining domains are not limited to information extraction but instead include automatic Email routing (e.g. automatically distribute Emails to the appropriate departments), spam filtering or fraud detection (e.g. unusual insurance claims are automatically routed for further investigation). [30]

3.4.4. Web Usage Mining

Web usage mining is about the behaviour of site visitors. Its goal is to show what users do on the site, how often they click on a specific menu button, in which order they follow certain links and so on. A famous and free to use example for usage mining software is Google analytics. It offers graphical representation of statistical values like the overall number of visitors, the way on which they get to a page (e.g. via feed, mail, through a paid link or by typing in the URL by hand), how long they stay there and how social networks are influencing the success of the site and company. Social networks can have influence when a user takes content from the company site and posts it or a link to it on a social platform. There he shares it with other users and initiates discussions about the company sites' content.



Besuchszeit auf Website nach Land

Land/Gebiet	Besuche	Durchschn. Besuchszeit auf Website
USA	67.445	00:01:54
Vereinigtes Königreich	18.948	00:01:37
Indien	8882	00:00:58
Kanada	6371	00:01:02
Deutschland	5845	00:00:32
Frankreich	5243	00:00:38

Figure 3.5: Google analytics offers graphical representations of website usage statistics[9]

Analysing web usage data collected in log files can help organisations to determine the value a customer generates for the company within his life as customer, it can help to develop cross-marketing strategies between products and services, evaluate the efficiency of marketing campaigns, optimize the functionality of web-based applications, offer a higher amount of user-personalized content and find the most effective logical structure for the company's web space.

Conducting usage mining requires access to server log files making it only suitable for website owners or administrators. For web mining on third party websites like it is done during the practical part of this thesis usage mining is not applicable.

[9] Source: `https://www.google.at/analytics/` (Mar 2014)

3.5. Online Communities

After it has been shown which possibilities exist to extract and process information gathered from the web we will now have a look at the data sources for the crawling and mining process. The sources for this thesis' work are online communities, also known as virtual communities or e-communities.

3.5.1. Concept definition

The term online community describes organized communication inside an electronic contact network. A group of individuals can meet on a common technical platform having or developing some kind of relationship to each other. The technical platform mostly is a webpage but can also be a mobile application or virtual environment. Communication takes place asynchronous and place independent. An online community can serve members in two ways: On one hand it can support information exchange between known and unknown participants while on the other hand it can manage user relationships. [32]

Online communities exist in many varieties. They can be as different as a health care platform that exists to help their members with serious illness and just-for-fun platforms with the only purpose to share funny pictures. However three characteristics have shown to be typical for all of them consistently in a large number of literature sources [33][34][35][36]:

- Members share common interests
- They have common experiences and needs
- Members evolve advantageous social relationships through which they get access to important resources, evolve strong interpersonal feelings of togetherness and feel a sense of common identity.

In 1996 a group of academics with different technical backgrounds held a workshop to identify a set of core characteristics of online communities that are very similar to the previously stated factors [28]:

- Members share a common interest, goal or need or belong to the community because of a common carried out activity.
- Members are continuously and actively contributing to the community. This often results in intensive interpersonal interactions, strong emotional relationships and common activities.
- Members have access to common resources. Accessing those resources underlies certain rules.
- Exchange of information, support and services among members is of high importance.
- There is a common context of social principles, speech and communication protocols.

3.6. Types of Online Communities

Several types of online communities exist which can be analysed with structure and content mining tools. The following online community type classification combines various classification attempts found in a bunch of literature. Slightly other arrangement of may be possible, but the general community types remain the same [37][1][38][39][40]:

- **Web rings** bring web sites about a specific topic together at one place. Users can add pages and comment pages already uploaded (example: WebRing©)
- **Product rating platform** on product experience report sites. These sites are usually coupled with a retailer and price comparison of the reviewed products (examples: Epinions, Dooyoo)

- **Mailing list servers** are used to supply a selection of members, who have subscribed to the mailing list, with new messages, newsletters or postings. It is usually a one-sided sort of community. The most common known usage of list servers is within newsletters.
- **Micro-blogging services** allow users to post very short messages on a community platform. Other members can read those messages and if they like it they may subscribe to the author to get notifications whenever new messages are posted from that author. (example: Twitter)
- **Online chat rooms** are websites, sub-pages or stand-alone applications such as messenger software (e.g. Yahoo Messenger, Skype or IRC) that allow real time communication between two or more participants. Normally this is done on textual basis, but it is often possible to support communication with video or audio. Chat rooms can either be open for any type of (small-)talk with the purpose of getting to know other people or they can be dedicated to a certain topic. Often people search for potential partners. Sometimes it is possible for users to open their own chat rooms either to make a private chat room where only selected user can participate or to open a topic related chat room. Chat rooms are often combined with other types of online communities such as forums or blogs. (examples: Friendscout24, chatroom, Superchat)
- **Media orientated communities** focus on user provided media rather than the person that uploaded it:
    - Wiki community where content is jointly created through several participants. (example: Wikipedia)
    - In voting or rating communities members are rated through other members. Rating criteria are mostly photos. (example: Hotornot)
    - Other media orientated communities whose members share content to show off something they have created or copied to a specific group of people or everyone using the portal. Usually the shown content can be rated and commented. (example: Youtube)
- **Topic oriented communities** or specialized information communities, unite people sharing a common interest. Users contribute to these platforms by posting personal experiences or help other users with problems they have posted. Mostly discussions in these communities are done through a forum that forms the centre of such a community site. Topic oriented communities include but are not limited to patient support, education or technical communities (e.g. Healthcentral, Javaranch, Whathifi). Some specialized versions of topic oriented communities can be distinguished:
    - Local online communities are a sub-form of topic orientated communities. They are dedicated to a certain place like a specific town usually letting only those people participate that are living in this town or place (examples: Dronefield Online).
    - Online research communities are topic orientated communities which are founded to support a group of students, scientists or other researchers by providing a platform where project participants can exchange information. Popular forms of these communities are e-learning platforms that support courses at universities (example: TUWEL)
    - Another special type of topic orientated communities is that who assembles a bunch of people to achieve a specific goal. This can be done to gather support for projects such as hindering companies to destroy nature or raising their voice to protest against political behaviour. After the goal is achieved, the community loses its purpose (examples: Save the arctic, Free pussy riot).

- **Commercial online communities** that are under supervision of a company are often integrated in an online shop. They include customer forums, chats, blogs and product ratings. (e.g. BattleNet Diablo 3 discussion page, Siemens employees blog, Amazon)
- **Virtual Worlds** are communities that take place in a computer generated fantasy world. The most famous type of such virtual worlds is the category of Massively Multiplayer Online Games (MMOG). They started out in 1991 with the most typical form MMORPG (Massively Multiplayer Online Role Playing Game), when AON released a game named Neverwinter Nights [41]. In that game every player could create his virtual alter ego – his avatar – with freely adjustable looks and skills. Up to 50 players could simultaneously walk around in a virtual environment, talk to virtual characters or to other players, collect or buy equipment and fight enemies. New equipment and experience gained through won fights made the character stronger and future fights easier. Most modern MMORPGs, although their look and sound has vastly improved and they offer much larger worlds as well as many more opportunities for a player to interact with that world, are still using the same basic game concept. It still motivates players by offering them ongoing small feelings of success with every better item they find and every mission they accomplish or fight they win that results in an experience gain and character improvement. The probably most widely known and still one of the most popular MMOGs these days is World of Warcraft, first released in 2004 with a maximum of 12 million active players in 2010 and 9.1 million players in the end of 2012 [42]. Besides MMORPGs there are two other types of MMOGs: Massive Multiplayer Online First Person Shooter (MMOFPS) and Massive Multiplayer Online Real Time Strategy (MMORTS). In MMOFPS players are fighting each other either alone or in teams on large battlefields. In MMORTS the player is not controlling just one character but has the authority over a whole army of units and buildings. The player can command his units to collect resources, construct buildings, attack an enemy or do other things. Besides interacting within the game players are able to chat or talk to each other. Sometimes they meet in online community websites or real life, either one on one or in groups. An example for a local online community that formed around the topic World of Warcraft with the purpose to meet teammates in real life is the NYC World of Warcraft Meetup Group (`www.meetup.com/nycWoW`).

Figure 3.6: MMORPGs now and then: Neverwinter Nights (left)[10] and World of Warcraft (right)[11]

- **Social network services** are platforms where users usually create profiles often containing personal data. On these profiles users can publish all sort of information relating to him. They can, for instance, state in which topics they are interested, post pictures or textual information about activities they have done or other private pictures or other media about themselves or

[10] Source: `http://www.bladekeep.com/gallery/displayimage.php?pid=204`

[11] Source: `http://wow.gamona.de/mists-of-pandaria/monch/`

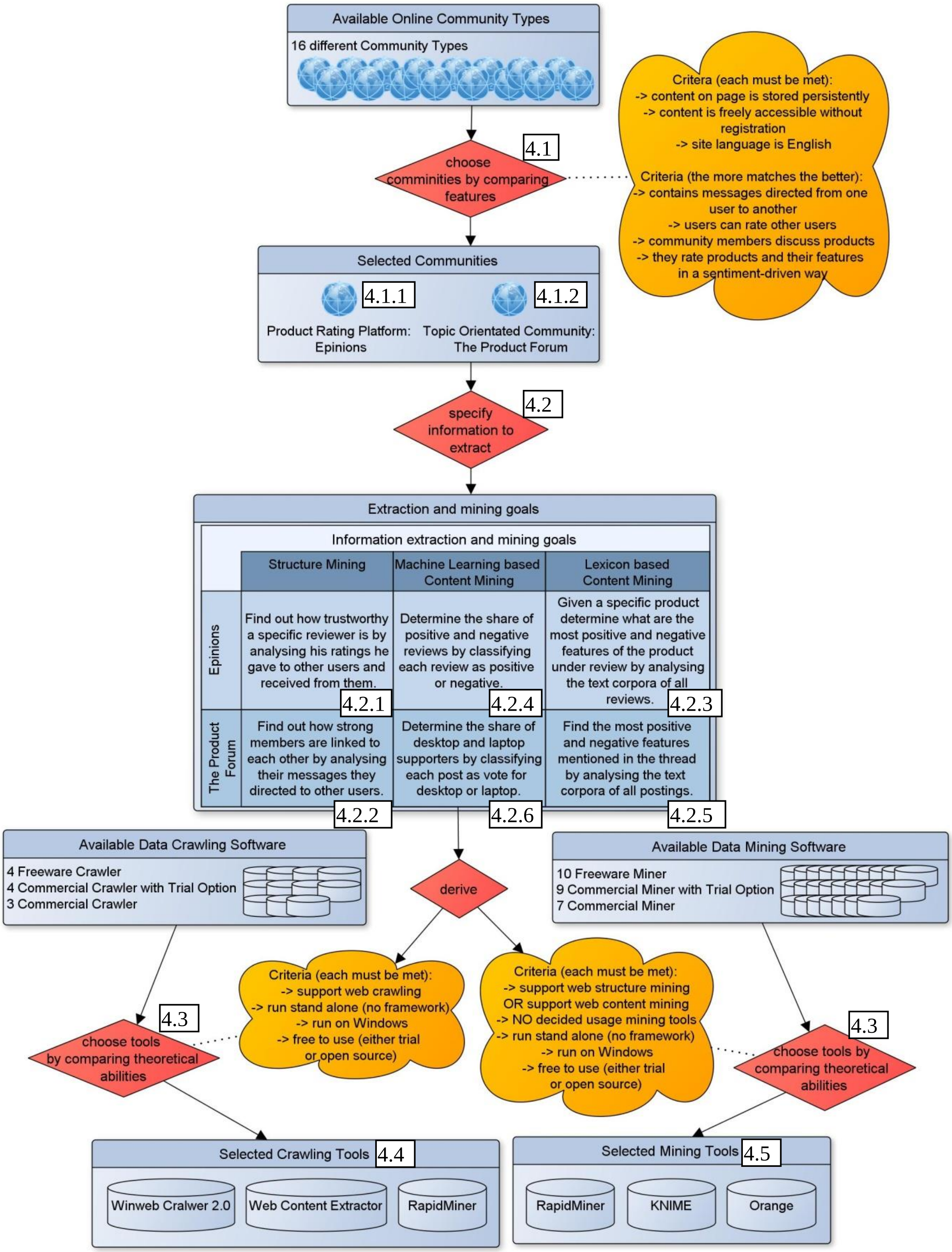
they can show up something that caught their attention and they want to show it to other members. Such platforms can be dedicated to a certain group of people (e.g. conservationists or a specific population group) or they can be generic. Generic social network services are often used to stay in contact with friends or colleagues, even if they live far away and could not keep in touch otherwise. These communities often offer the possibilities to create the presence of a commercial entity. E.g. there exists a page on Facebook that is created by Austria's largest computer reseller DiTech. It is used to promote special offers and computer related news. Customers can use the site to post their opinions about offers, products or new developments, use the site to post complaints or ask for technical support. (examples: Facebook, Myspace, Stayfriends, Flickr)

- **Public blogs** can be used by a person to periodically express his opinion about a specific topic or make an announcement in form of a short article. Normally readers are allowed to leave comments. Besides text based blogs there are media blogs where the blogger posts video, audio or picture material. (example: Hubspot)

Although in most cases communities clearly belong to a certain type stated above, it can be the case that a community fits into several categories. For instance YouTube combines a media orientated community (uploaded videos) with social network services (every user has his own profile that can be personalized by himself as well as commented and followed by other users). Another case is the BattleNet webpage, which is a perfect example for a commercial online community but features public blogs as well.

4. Selecting appropriate Communities and Tools

This chapter discusses the first step of the practical data mining process: the selection of crawling tools, mining tools as well as the communities these tools will be used on. Three selection processes will be done as it is shown in the overview graph in figure 4.1.



Available Online Community Types
16 different Community Types
4.1
choose comminities by comparing features
Critera (each must be met):
-> content on page is stored persistently
-> content is freely accessible without registration
-> site language is English
Criteria (the more matches the better):
-> contains messages directed from one user to another
-> users can rate other users
-> community members discuss products
-> they rate products and their features in a sentiment-driven way
Selected Communities
4.1.1
4.1.2
Product Rating Platform: Epinions
Topic Orientated Community: The Product Forum
4.2
specify information to extract
Extraction and mining goals
Information extraction and mining goals

	Structure Mining	Machine Learning based Content Mining	Lexicon based Content Mining
Epinions	Find out how trustworthy a specific reviewer is by analysing his ratings he gave to other users and received from them. 4.2.1	Determine the share of positive and negative reviews by classifying each review as positive or negative. 4.2.4	Given a specific product determine what are the most positive and negative features of the product under review by analysing the text corpora of all reviews. 4.2.3
The Product Forum	Find out how strong members are linked to each other by analysing their messages they directed to other users. 4.2.2	Determine the share of desktop and laptop supporters by classifying each post as vote for desktop or laptop. 4.2.6	Find the most positive and negative features mentioned in the thread by analysing the text corpora of all postings. 4.2.5

Available Data Crawling Software
4 Freeware Crawler
4 Commercial Crawler with Trial Option
3 Commercial Crawler
derive
Available Data Mining Software
10 Freeware Miner
9 Commercial Miner with Trial Option
7 Commercial Miner
4.3
choose tools by comparing theoretical abilities
Criteria (each must be met):
-> support web crawling
-> run stand alone (no framework)
-> run on Windows
-> free to use (either trial or open source)
Criteria (each must be met):
-> support web structure mining OR support web content mining
-> NO decided usage mining tools
-> run stand alone (no framework)
-> run on Windows
-> free to use (either trial or open source)
4.3
choose tools by comparing theoretical abilities
Selected Crawling Tools
4.4
Winweb Cralwer 2.0
Web Content Extractor
RapidMiner
Selected Mining Tools
4.5
RapidMiner
KNIME
Orange

Figure 4.1: Community and tool selection process overview. Each number in a white rectangle describes the chapters where the corresponding task or result is described in further detail.

4.1. Selecting Communities suitable for Analysis

Since this thesis cannot outline all sorts of communities presented in the previous chapter, those most appropriate for web mining have to be identified and selected.

Usually online research follows a goal that defines what information should be gathered during the research process. This work does not demand to gather some specific information but instead focusses on the information gathering process itself. Demanded information therefore will be defined in a more common way first to determine which type of community page is appropriate as data source and then after a specific page has been chosen it will be specialized to a specific topic.

Factors that make an online community suitable for answering a specific question are [1]:

- The community is focused on the relevant topic
- It should have as much traffic as possible
- It contains a multitude of discrete message posters
- User posted, topic-related data should be rather detailed
- Interaction between members is done in a way required by the research question

In addition first tests and literature have shown there are some more preferable properties a community should have to be suitable as web mining data source [1]:

- Information should be persistent meaning information a user posted in the past should be permanently stored and made available for public access and review. Online Chat rooms, list servers as well as virtual worlds sometimes use logs, but they usually are not freely accessible for someone other than the communication participants and therefore those communities do not meet that criterion.
- Information should be freely available without registration. Although data mining can be executed in environments that need registration to view user content as shown in [19], it is not possible without getting allowance from those users that are going to be analysed. This makes the mining process much more complicated. Social network and micro blogging – although very popular – do not meet this criterion.
- Since many tools make use of built-in English word-recognition dictionaries, results on German webpages can be disappointingly bad in many cases. These tools, for example, cannot distinguish names of persons, companies or places from other word types since they cannot recognize these words in their dictionary. Although it is possible to include external dictionaries in many cases, most existing sentiment dictionaries are in English too. In order to maximize compatibility, reviews of sites that use other languages than English should be avoided.

Considering these factors and restrictions there still remain many community types like web rings, blogs, topic related communities, media orientated communities and rating communities. To figure out which one will be best, we have to take a deeper look in what market researchers are seeking in those communities.

But why does this thesis put its focus on answering market research questions? There are other fields of interest one could explore like predicting financial trends from text news wires, spam filtering, identifying fraudulent transactions or automatic building of ontologies (for subject classification). [43] Nonetheless the application of conducting marketing as well as gaining knowledge from and about customer groups is possible in many ways making up a significant part of the whole application list. Applications include market segmentation, determining customer churn, direct marketing (e.g. identify best prospects for marketing broadcasts), interactive marketing (show each individual site visitor content he is most likely interested in), market basket analysis (what products and services are commonly purchased together?) or customer behaviour trend analysis. [43]

In recent years customer-made reviews were getting more and more important. According to a study conducted by Myles Anderson in June 2013 over 50% read customer reviews to determine the quality of a local business (see figure 4.2). Over 70% are affected in a positive way by reading positive consumer reviews. Impressing 79% trust online customer reviews as much as personal recommendations of friends at least for some businesses. [44] While that study focusses on local businesses like restaurants, doctors or bars, this thesis covers product reviews. Numbers may differ but a similar ascending trend may be assumed for this category as well.

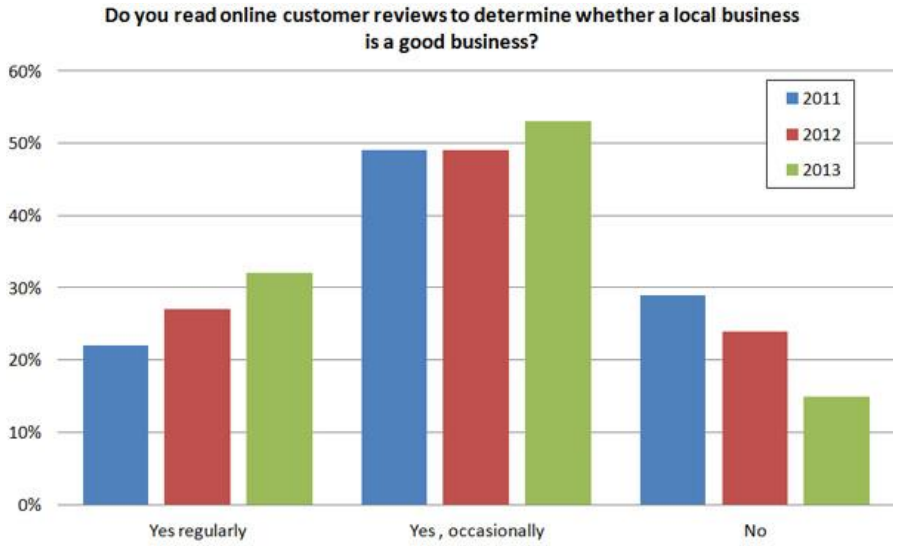


Figure 4.2: How many customers read online reviews to build their opinion? [44]

Marketing experts generally agree that one main factor for a successful business is to know your customer target group [45][46]. What web structure mining can do for market researchers is to discover links between users of online communities. If you know more about people who wrote reviews about your product or who released an opinion about a person (e.g. somebody belonging to your company), with whom those reviewers interact, to what other sites, products or persons they refer or where they have published other content, you will get a deeper insight into characteristics of those people, add context to what they wrote and maybe explore new links to other topics or persons of interest.

Sounds good, but how can web content mining in online communities actually be used for market research purposes?

Related work by Mr. Morzy has shown that it is possible to use structure mining to verify the trustworthiness of ratings by analysing the social environment inside the community. E.g. seller A on Ebay that got 100 positive ratings from buyers that have made only few or no business transactions with other buyers from A can be regarded as unobtrusive and most likely a regular seller with regular ratings. If seller B got 100 positive ratings from only 10 different persons, the price of a large share of ratings is only 1 Euro and 8 of those 10 buyers made a large number of business transactions with each other, it is very likely that seller B cooperates with those buyers to get a higher rating in an unauthorized way.

Web content mining on the other hand can be used to automatically distinguish positive from negative ratings on a product rating site and generate a summary of positive or negative product features. Experimental work done in earlier publications has shown that it is possible to specify positive or negative terms for a specific domain such as hotels or cameras and use this term list to get a rather detailed line-up of negative or positive product attributes [47][23]. This adds a lot of additional value to reviews since you are no longer restricted in seeing the average overall score (in Epinions it is some value on a scale from 1 to 5 points) but instead get a list of what exactly the positive and negative rated attributes are. For instance, a camera can have a positive attribute “good picture quality” and a bad attribute “short battery runtime”. This kind of mining was discussed under the term sentiment analysis in chapter 3.4.3.

Regarding the considerations above as much of the following criteria as possible should be met in order for a community to be appropriate as data source for the tool test phase in the following chapters. Not all of these criteria have necessarily to be met but the more the better:

- The community members make use of direct communication among each other meaning it can clearly be determined when one user directs his message to another user.
- Community members can rate other members and can be rated by other members as well.
- Community members take part in discussions about products – regardless, what type or label the product is.
- Community members rate products and product features in a clearly sentiment-driven way.

After taking these requirements into account most community types turn out to be inappropriate for the practical part as can be seen in the evaluation tables on the following pages.

	Web Rings	**Product rating platform**	**Mailing list servers**	**Micro blogging services**
Example pages	http://hub.webring.org/	http://www.epinions.com/?sb=1 http://www.dooyoo.co.uk/	http://www.absorber.se/newsletter/newsletter_online	www.twitter.com
Persistent on page	Yes	Yes	Yes	Yes
Freely accessible	Yes	Yes	(mostly) No	No
English language	Yes	Yes	Yes	Yes
Structure mining:				
User rating system	No	Yes	No	No
User contact/ interaction network	Yes (through comments)	Yes (through comments)	No	Yes (followers)
Content mining:				
User based product rating	No	Yes	No	No
User based product discussion	No	Yes (through comments)	No	No
Qualifies for test phase	No	Yes	No	No

Table 4.1: Community evaluation table: Testing which communities fulfil the requirements as data mining source for practical tool tests – part 1.

	Online chat rooms	**Media orientated Wiki**	**Media orientated voting**	**Other media orientated community**
Example pages	http://www.friendscout24.at/-z/de_AT/lexikon/chatroom.html http://www.superchat.at/	www.wikipedia.org	www.hotornot.com	www.youtube.com
Persistent on page	(mostly) No	Yes	Yes	Yes
Freely accessible	No	Yes	No	Yes
English language	Yes	Yes	Yes	Yes
Structure mining:				
User rating system	No	No	Yes	No
User contact/ interaction network	No	No	No	Yes (abonnements)
Content mining:				
User based product rating	No	No	No	Yes (some media represent product reviews)
User based product discussion	No	No	No	Yes (as comments to product media)
Qualifies for test phase	No	No	No	No (media cannot be extracted automatically)

Table 4.2: Community evaluation table: Testing which communities fulfil the requirements as data mining source for practical tool tests – part 2.

	Topic oriented communities	Local online communities	Online research communities	Goal directed communities
Example pages	http://www.healthcentral.com http://www.javaranch.com/ http://www.theproductforum.com	http://www.dronfieldonline.co.uk/going-out-in-dronfield-c631.html	https://tuwel.tuwien.ac.at/	http://rising.savethearctic.org/arctic_rising/ http://freepussyriot.org/
Persistent on page	Yes	Yes	Yes (limited)	Yes (limited)
Freely accessible	Yes	Yes	No	Yes
English language	Yes	Yes	Yes	Yes
Structure mining:				
User rating system	No (but usually the number of user posts are displayed)	No	No	No
User contact/ interaction network	Yes (forum post answers)	Yes (forum post answers)	Yes (forum post answers)	No
Content mining:				
User based product rating	Yes (sometimes threads are reviewing a product)	No	No (in exceptional cases maybe yes)	No
User based product discussion	Yes	No	No	No
Qualifies for test phase	Yes	No	No	No

Table 4.3: Community evaluation table: Testing which communities fulfil the requirements as data mining source for practical tool tests – part 3.

	Commercial online communities	Virtual worlds	Social network services	Public blogs
Example pages	http://eu.battle.net/d3/en/ http://sispsebr.wordpress.com/ www.amazon.com	http://www.freebrowsergamer.com/2008/12/best-browser-based-mmorpg.html	www.facebook.com http://www.myspace.com http://www.stayfriends.at/	http://blog.hubspot.com
persistent on page	Yes	No	Yes	Yes
Freely accessible	sometimes Yes	No (ingame messages are not visible to others)	No	Yes
English language	Yes	Yes	Yes	Yes
Structure mining:				
User rating system	sometimes Yes	No	No (except if friendships are counted as votes)	No
User contact/ interaction network	sometimes Yes	No	Yes	Yes (through comments)
Content mining:				
User based product rating	sometimes Yes	No	Yes (in postings and comments)	Yes (in blog posts and comments)
User based product discussion	Sometimes Yes	No	Yes (postings and comments)	Yes (in comments)
Qualifies for test phase	some forms Yes (e.g. Amazon)	No	No	Yes

Table 4.4: Community evaluation table: Testing which communities fulfil the requirements as data mining source for practical tool tests – part 4.

As the table shows only product rating platforms and commercial online communities fulfil all of the requirements. Research work so far as mentioned in the related work chapter has shown that acceptable results can be obtained by conducting structure mining on user ratings in online auctions. It can be assumed that structure mining on rating communities will lead to good results as well. On the other hand content mining could be conducted on a rating community in [47] as well giving evidence to the selection done within this chapter.

Commercial online communities can be any form of community that is run by a company with commercial interest. This means commercial communities are not clearly distinct from other community types but instead just add the business factor to them. For instance, the commercial community Amazon is feature-wise very similar to the non-commercial product rating community Epinions providing an almost identical rating and user recommending system. Both support writing reviews, mark them as helpful or not and vote for users that generally write high quality reviews. Since commercial communities offer the same functionality as their non-commercial pendants, they will be excluded from the test phase.

With still six out of seven criteria fulfilled topic oriented communities and public blogs are two options worth considering. While public blogs are rather focussed on the blog statement and discussion between users plays only a minor role, on product based communities those discussions are much more important. Mostly community members on such platforms ask for help in an upcoming decision for or against a product or want to exchange and discuss their experience with certain products. In blogs authors normally just want to express their own opinions and put less emphasis on discussion and interaction with other users. This is the reason why the second type of community under review is a content based community rather than a blog.

Speaking of content based communities, already published publications have shown that a link analysis between forum members can be conducted using the forum member names that are mentioned in posts. By counting the number of times a member mentions another member in his posts, a link analysing graph can be deduced. Content mining on the other hand can be done in the same way as on rating platforms since the basic pattern of information representation remains the same: Each user writes text in a field and fields are displayed among each other. The main difference compared to ratings is that forum threads usually contain much more noisy data like posts that are off topic or sentiment-neutral and will likely affect content recognition results.

4.1.1. Rating Community: Epinions

After selecting the community type it is necessary to select specific community sites. As rating community Epinions (`www.epinions.com`) was selected. Epinions is a general consumer review site for all sorts of products reaching from cars and computers to sport equipment and hotels. It was founded in 1999 and is owned by Ebay since they acquired it in 2005. According to URLSPY[12] the site has about 2,4 million unique visitors per month and an estimated worth of 1,4 million US$ making it the 3500th most popular website worldwide in February 2014. Other communities such as Amazon.com or TripAdvisor.com would be as appropriate as Epinions in terms of review presentation making content crawling and mining as easy on all of these sites. Epinions features a more advanced user trust system clearly displaying which users trust a certain member. This makes visualizing links possible like it has been done in online auctions where you see who has rated a member positive or negative [15]. In contrast Amazon.com only features an anonymous trust system where users can

[12] See `http://au.urlspy.co/www.epinions.com`

collect “helpful” votes for their reviews without knowing who has rated them and TripAdvisor features mostly the same anonymous system.

4.1.2. Topic Oriented Community: The Product Forum

The chosen topic oriented community was a product discussion forum named “The product forum” (`www.theproductforum.com`). It is not the largest topic oriented community with 2540 members, 2335 threads and 13298 posts in February 2014[13]. The largest community forums out there have up to 1,8 billion posts and 23 million registered users (largest site on February 2014: Gaia Online)[14]. Most posts on The Product Forum have less than a few hundred words and threads usually would not exceed 10 posts. Since many commercial tools with trial options that will be tested afterwards have a limited result size limit for their free trial versions, a smaller data set size helps not exceeding data set limits and makes results more comparable to freeware tools without such limits. Despite its relative small size, the forum provides a well-balanced set of features like video reviews, blogs and a forum. Its structure allows clearly arranged analysis and makes results manually easily revisable.

4.2. Specify Information that should be extracted from Communities

A distinction between structure and content mining has to be made in order to consider different types of information to gather. For each selected community one structure mining and two content mining tasks will be defined summing up to six mining tasks.

4.2.1. Structure Mining on Epinions

Using mining tools under review the following question will be answered for Epinions:

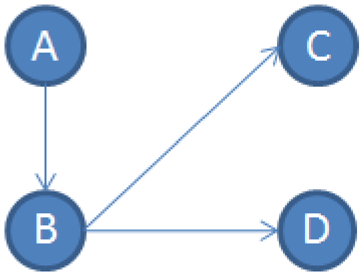
Given a specific review about a product – how trustworthy is the reviewer? Is he given trust by independent, trustworthy members?

The site has implemented a function that enables users to express trust to another user by clicking on the “trust” button. It has been shown, that such trust systems are not fully reliable. Trust points can be received by friends only and vice versa can be given to friends or partners regardless of the actual review quality. In online auction seller reviews there exist a remarkable percentage of fraudulent ratings that can be partially exposed by exploring how reviewers are linked to each other [15]. If those people witnessing their trust to the reviewer are suspiciously strong linked to each other giving plentiful trust points to each other but less to third persons, the reviewer and his review should be considered less trustworthy.

Expected result: Each tool will be used in a way to output a link graph showing who has given trust points to the reviewer and how they are linked to each other. Since the main goal of this thesis is to show the tools’ capabilities and not to analyse the results, it is not necessary to find clear evidence for or against suspicious trust rating behaviour. The result should accurately represent reality having a form similar to that shown in figure 4.3.

[13] `http://www.theproductforum.com/forums/forum.php`

[14] `http://www.toptenz.net/top-10-biggest-internet-communities.php`



User 1	User 2
A	B
B	C
B	D

Figure 4.3: Structure mining result format. The graph on the left shows the user structure graph while the table next to it shows the underlying node data format.

4.2.2. Structure Mining on The Product Forum

Speaking for the product forum, the structure mining tools should be able to answer following question:

How strong are forum members linked to each other? Who communicates with whom and to what extent?

The underlying analysis can be done by recognizing user names and searching for post messages that include those names. If a name is mentioned in a message, it is likely addressed to this user and will be counted as communication link from the author to the user mentioned in the message.

Expected result: The expected result should be a directed graph which shows how often users have mentioned other users in their posts. The graph and its underlying data format will have the same format as the structure mining graph from Epinions shown in figure 4.3.

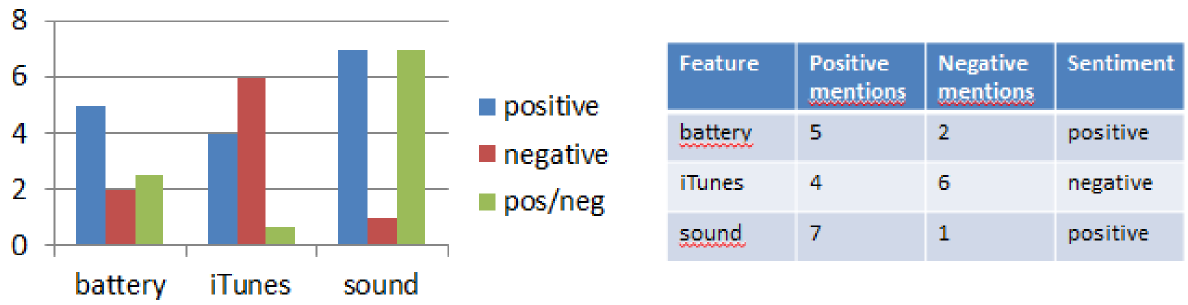
4.2.3. Lexicon based Content Mining on Epinions

The question that shall be answered during the lexicon based content data mining procedure is the following:

For a given product – what are the positive and negative features that can be derived from the product reviews?

To answer this question one product with a large number of review texts will be chosen on the Epinions rating page. A product with a remarkable amount of 159 reviews at the moment this thesis is written is the Apple iPod Video 5th Generation White (30 GB) MP3 Player. This should be more than sufficient to get comprehensive results.

Expected result: A list of features marked as positive and negative with observed frequency of occurrence. Higher occurrence means this feature is mentioned more often and therefore has higher relevance.



Feature	Positive mentions	Negative mentions	Sentiment
battery	5	2	positive
iTunes	4	6	negative
sound	7	1	positive

Figure 4.4: Expected result example for content mining on Epinions. Within the table on the right are the extracted feature mentions as well as how often they are mentioned in a positive or negative manner. On the left this data table is visualized.

4.2.4. Machine Learning based Content Mining on Epinions

To test text classification abilities, following question will be answered:

Which reviews are overall positive and which ones are negative? What is the overall share of positive or negative reviews?

Despite this type of information is already there on the Epinions website reviews because of the predefined field “recommended” that is marked as yes or no by the user, there are some websites out there that does not have that feature like the UCI-Cinema website (`http://www.uci-kinowelt.at/`). This website offers the possibility for visitors to rate a film on a scale from 1 to 7 points on one hand and to independently write a short review about the film on the other hand. That is why the mining tools will gather already known information in this case that could easily applied to other sites that lack this information. To do so a learning model will be trained with varying sizes of training sets that are review text corpora.

Expected result: The outcome will be a classified list of reviews each one either marked as positive or negative as well as a percentage of positive and negative reviews.

4.2.5. Lexicon based Content Mining on The Product Forum

As with structure mining, a second task will have to be performed by the reviewed tools. In this case the tools will be used to pick up a single forum thread from “The Product Forum”. In the category “Desktops, Laptops, Notebooks & Consoles”, the thread with most posts (42 at the moment this thesis is written) in it is named “Desktop vs. Laptop?” As the name suggests it houses a discussion about the pros, cons of desktop and laptop computers as well as the preferences of the discussion participants. The question that the content mining procedure should assist to answer is following:

What are the advantages and disadvantages of laptops or desktops regarding the posts in this thread?

Expected result: The mining procedure should deliver data similar to the content mining task conducted on Epinions. Features coupled with positive and negative opinions should be identified and presented in a list just as shown in figure 4.4 in the previous subchapter.

4.2.6. Machine Learning based Content Mining on The Product Forum

As with content mining on Epinions a text classification based question will be asked as well:

What is the overall share of laptop and desktop supporters?

Again a learning model will be trained with the whole set of posts labelled manually used as training set.

Expected result: The result will be a classified list of posts as well as a percentage of Laptop and Desktop supporters.

4.3. Data Crawling and Mining Tools

In the previous chapter six mining questions have been specified. To successfully answer them the data extraction part has to be done in the next step. A bunch of free software tools, including both crawler and miner, will be installed and tested for the ability to generate results that possibly match the expected results defined in chapter 4.2. Depending on whether the software under review is a crawler, a miner or both in one product, the different abilities and test results of each tool will be compared and inserted in a comparison table.

Since the number of web crawling and data mining tools available on the market is unmanageable large it is impossible to test every tool within the scope of this thesis. Instead selection according to specific criteria has to be done to find those tools most suitable for the desired tasks before actual tests are run.

4.3.1. Selecting appropriate Mining Tools

Seeking for mining tools was done using various search engines as well as including tools mentioned in related work. Search has resulted in finding a couple of crawling and mining tools that differ in terms of system requirements, data input and output formats, crawling and analysing functionality such as text and structure mining capabilities and the way they can be executed. For instance, they can be stand-alone programs or just toolkits that can be used to write one’s own mining tasks in program languages as e.g. Java. To make research as straightforward as possible, all mining tools included in the comparison should meet the following criterions:

1) Web usage mining depends on server log files inaccessible to anyone else than the website administrator or authorized people. While web usage mining can give valuable insights into customer and visitor behaviour, it is not suitable for the desired content and structure mining tasks. Web usage mining functionality is therefore ignored and tools supporting web usage mining only will be excluded from further testing.
2) They have to be ready to run out of the box requiring no programming besides rather simple scripts within the program itself. This excludes frameworks or code fragments since this thesis will not focus on programming but only on appliance.
3) They can be run on Windows operating system. In first tests some Linux-only tools like ASPseek have shown to be rather complicated to install and to use making them less user friendly and convenient to use. Since user friendliness is a rather important quality measure for software and most tools – especially more sophisticated tools with most extensive features – run on Windows anyway, software that is limited to Linux or other free OS will be excluded.

4) Most important is the programs ability to fulfil the requirements to answer the questions stated in chapter 4.2. This requires either structural mining abilities that are able to link user names on the web or content mining abilities that offer word and phrase analysis or both of them. Another functionality that is requested is the ability to crawl websites. Tools that offer an additional crawling functionality or completely independent crawling tools will be included in the test phase as well while tools that obviously are not capable of performing at least one of these tasks will be excluded.
5) They have to be freely available. Since this thesis lacks a sponsor financial efforts have to be kept within narrow confines. Tools require offering at least a free test version in order to qualify for practical testing.

Keeping these restrictions in mind, tools will be selected within a process of elimination in the following sub-chapter. An overview table is created where the tools are tested against required features defined above. After pre-selecting a set of tools based on these four factors, the remaining tools are going to be installed and used in chapters 5 and 6 to answer questions defined as accurate as possible.

4.3.2. Available Open Source and Freeware Tools

This category includes solely those tools that can be used to their full extent completely free of charge.

	Easy.Data.Fox	Keyword Crawler	KNIME
Developer/ Publisher	Easy.Data.Mining	WebKeySoft	KNIME GmbH
Homepage	`www.easydatamining.com`	`http://keyword-crawler.software.informer.com/`	`http://www.knime.org/`
Short description	This freely available tool from the German company Easy.Data.Mining is a very simple to use data prediction tool for numerical data stored in .csv tables.	Very simple to use tool, that mainly focusses on crawling and analysing websites to discover broken links or display a list of most frequently used words on the webpage under review.	Very comprehensive data mining tool based on Eclipse (www.eclipse.org) with many, partially community created extensions that offers almost any type of data preprocessing, analyzing, predicting, classifying, visualizing of numeric and textual data.
Import formats	.csv tables	html pages	tables, text files, various Databases (e.g. SQL, Java), ARFF[15], PMML[16]
Output formats	graph	list of words	table, fext file, graph, database entries, ARFF, PMML
Numerical[17] data forecasting or model learning	Yes	No	Yes
(Past) numerical data pattern analysis	No	No	Yes
Runs stand-alone	Yes	Yes	Yes
Runs on Windows	Yes	Yes	Yes
Supports (Web) structure mining	No	No	Yes
Supports text mining	No	Yes but word count only	Yes
Supports web crawling	No	Yes but limited to single words	No
Qualifies for the analysis goal of this thesis	No (lacks desired crawling and text mining function)	No (too few text mining and crawling options - can only retrieve frequent word count)	Yes

Table 4.5: Tool overview comparison table for open source software

[15] Attribute Relation File Format (defined by WEKA – see its description on next page)

[16] Predictive Model Markup Language: based on XML, exchange format between different data mining programs

[17] Including text fields with limited possible values (e.g. credit rating can be AAA, AA, A, B or C. Those values are processable just as a number e.g. from 1 to 5)

	WEKA	Orange	jHepWork
Developer/ Publisher	The University of WAIKATO	solely community created	Dr. Sergei Chekanov
Homepage	`http://www.cs.waikato.ac.nz/~ml/weka/`	`http://orange.biolab.si/`	`http://jwork.org/jhepwork/`
Short description	WEKA is a collection of machine learning algorithms for data mining tasks written in java. They deliver functionality for preprocessing, classification, regression, clustering, association rules and visualization. WEKA offers a GUI to use the algorithms without programming skills (Weka Knowledge Explorer). KNIME, Orange and RapidMiner offer an implementation of these algorithms as well.	Very comprehensive data mining tool similar to KNIME but programmed in Python. It too offers a lot of numerical and text data analytic options that can be extended by downloadable and self-programmable widgets.	jHepWork is a powerful Java library including a programming environment that supports mining of numerical data, image manipulation and text processing.
Import formats	database, URL, .csv table, ARFF(recommended), java.io	text files, ARFF, xml, database, other learning model files	java.io, SQL, object based DB, XML, C++, Google Protocol Buffers library
Output formats	graph, ARFF, java.io	graph	java.io, SQL, object based DB, XML, C++, Google Protocol Buffers library
Numerical data forecasting or model learning	Yes	Yes	No
(Past) numerical data pattern analysis	Yes	Yes	Yes
Runs stand-alone	Yes	Yes	No
Runs on Windows	Yes	Yes	Yes
Supports (Web) structure mining	No	Yes	No
Supports text mining	Yes but document classification only	Yes	Yes
Supports web crawling	No	No	No
Qualifies for the analysis goal of this thesis	No (too few text mining options, algorithms are implemented in other programs as well making testing this software obsolete)	Yes	No (would not run stand-alone)

Table 4.6: Tool overview comparison table for open source software

	ASPseek	**Tableau Public**	**Datapark Search**
Developer/ Publisher	Swsoft	Tableau Software	Maxim Zakharov
Homepage	`http://www.aspseek.org`	`http://www.tableausoftware.com/`	`http://www.dataparksearch.org/`
Short description	Aspseek is an internet search engine developed under the GNU GPL license that stores search results in SQL tables. It may be used as crawler.	Tableau public is aimed at rather unexperienced users that get fast graphical representations of mainly numerical or discrete values. You just have to open the file or database, choose from a number of graphical representations (bar, heatmap, scatter plot, pie chart, etc.) and then publish it online.	Datapark Search is a Web search engine for searches within a website or groups of websites. It runs on Linux, BSD, Solaris and CentOS and needs MySQL community server or ORACLE in order to run.
Import formats	Web sites only.	Access Database, Excel Table, Text file, SQL server, PostgreSQL, IBM DB2, various other server database formats	Web sites and local networks.
Output formats	MySQL databases	publish results online	MySQL or ORACLE database entries
Numerical data forecasting or model learning	No	No	No
(Past) numerical data pattern analysis	No	Yes	No
Runs stand-alone	Yes	Yes	No
Runs on Windows	No	Yes	No
Supports (Web) structure mining	No	No	No
Supports text mining	No	No	No
Supports web crawling	Yes	No	Yes
Qualifies for the analysis goal of this thesis	No (would not run under windows)	No (is not designed for text mining or structure mining tasks but for numerical and discrete data analysis. It is particularly strong in its heatmap-view abilities)	No (would not run stand-alone, would not run on Windows)

Table 4.7: Tool overview comparison table for open source software

	Arachnode	ADaM
Developer/ Publisher	arachnode.net LLC	IT and Systems Center, University of Alabama
Homepage	`www.arachnode.net`	`http://datamining.itsc.uah.edu/adam/`
Short description	Arachnode is a web crawler written in C# similar to a search machine running on SQL databases. In order to run it you have to install SQL Server 2005/2008, SQL Management Studio and Visual Studio 2008/2010 (Express). Besides an automated content extraction it offers the text mining capabilities text cleaning and bayes classification.	Toolkit consisting of over 100 components that can be used for data mining and image classification. Each component consists of an API like C, C++, etc., an executable file for supporting scripting tools like Perl, Python, Shell, and optional a web service interface for supporting web- or grid applications. Operations can run stand-alone making it ideal for parallel processing and grid computing. It does not support visualization. It is used by the NASA and National Science Foundation.
Import formats	user search rules, SQL Server Databases, .doc, .pdf, .ppt, .xsl files, HTML pages, EXIF[18] Metadata	Binary Image Format, gif, ARFF
Output formats	XML, XHTML, SQL database entries	Binary Image Format, gif, ARFF
Numerical data forecasting or model learning	No	Yes
(Past) numerical data pattern analysis	No	Yes
Runs stand-alone	No	No
Runs on Windows	Yes	Yes
Supports (Web) structure mining	No	Yes
Supports text mining	Yes	Yes
Supports web crawling	Yes	No
Qualifies for the analysis goal of this thesis	No (would not run stand-alone, needs SQL Management Studio and Visual Studio)	No (it is a toolkit)

Table 4.8: Tool overview comparison table for open source software

[18] Exchangeable Image File Format

	Databionic ESOM[19] Tools	**SharperLight**
Developer/ Publisher	Databionics Research Group, University of Marburg	Philight Software
Homepage	`http://databionic-esom.sourceforge.net/`	`http://sharperlight.com`
Short description	This program is designed to automatically organize multivariate data points into homogeneous groups. It breaks down a high dimensional space in a low dimensional one while preserving its topology. It can be used for unsupervised clustering and supervised classification. An ESOM is a SOM for at least a few thousand data points.	This back end application infrastructure framework allows users to merge data from various sources like Excel tables, data base entries, documents, images as well as financial and operational applications without the need of data migration. It requires MS SQL Server or Express Server to run. Implementations can be ordered on demand by contacting the developers.
Import formats	Tab separated values	Excel Table, data base entries (Access, MS SQL, DB2, Oracle, MYSQL, etc.), images, documents, SaaS[20] sources
Output formats	Tab separated values, graphical representation of SOM	Excel, Web Gadget, BIRT[21], SSRS[22], …
Numerical data forecasting or model learning	No	No
(Past) numerical data pattern analysis	Yes	Yes
Runs stand-alone	Yes	Yes but SQL Server is needed
Runs on Windows	Yes	Yes
Supports (Web) structure mining	No	No
Supports text mining	No	No
Supports web crawling	No	No
Qualifies for the analysis goal of this thesis	No (does not support web structure or content mining)	No (does not support web structure or content mining)

Table 4.9: Tool overview comparison table for open source software

[19] Emergent Self Organizing Maps: Development of high level structures through concurrence of several elementary processes in an independently running system.

[20] Software as a Service: Application and IT infrastructure are run at an external IT service provider and are used by the customer as an online service (mainly through a browser)

[21] Business Intelligence and Reporting Tools: An Eclipse-based open source project delivering business intelligence and reporting functionality for rich clients and web applications. `http://www.eclipse.org/birt/phoenix/`

[22] SQL Server Reporting Services tools and programming features for creating, deploying and managing reports as graphics, tables or freely defined forms. `http://msdn.microsoft.com/en-us/library/ms159106.aspx`

4.3.3. Available Commercial Tools with Trial Option

The category of commercial tools that offer a demo version still offers the possibility to use their product free of charge but only in a limited way. Many publishers limit the time the software works before requiring a registration fee. Alternatively or in addition to the time limit the size of data sets analysable by the software is often limited as well.

	perSimplex Divide and Impera	**Analysis Studio**	**VisuMap**
Developer/ Publisher	umaa, ltd.	Appricon Ltd.	VisuMap Technologies Inc.
Homepage	`http://www.persimplex.biz`	`http://appricon.com/`	`http://www.visumap.net/`
Short description	Tool for analysing and visualizing structured Data in .csv Files. Similar to tableau public and Data Applied, but offers other data mining functions and algorithms. The tools' unique feature is the ability to cluster data sets with similar trends. E.g. for a number of salespersons with known sales figures at specific points in time it is able to make one cluster for all salespersons with rising sales figures, a second with decreasing sales figures and a third containing those whose sales remained stable.	This data mining application can be used mainly for numerical analysis, to build product demand curves, predict sales levels, show relationships and correlation between multiple variables and other production and profit optimization purposes. It is not suitable for analysing text content.	Data visualization tool for data stored in tables that shows data in different 2D or 3D views. It allows data mapping, dimensionality reduction and clustering. It can be used for financial or market analysis.
Import formats	.csv	Excel	.csv, .txt
Output formats	graph	Excel, PDF, XML, csv	dynamically linked 2D and 3D graphs
Numerical data forecasting or model learning	No	Yes	No
(Past) numerical data pattern analysis	Yes	Yes	Yes
Runs stand-alone	Yes	Yes	Yes
Runs on Windows	Yes	Yes	Yes
Supports (Web) structure mining	No	No	No
Supports text mining	No	No	No
Supports Web crawling	No	No	No
Trial restrictions	maximum of 100 rows and 24 columns (= 2400 data values)	30 days test period	14 days trial period, <400 data values
Commercial version cost	for 1/2 years: <50.000 data values 330$/465$, <500.000 values 960$/1270$, <2.500.000 values 1810$/2320$	<5.000 entries 29,95$, <100.000 entries 49,95$, unlimited entries 599,95$	<400 data values 150$, unlimited data values 1.400$; unlimited values 1 year 700$; academic license 500$
Qualifies for the analysis goal of this thesis	No (only suitable for numeric variables and predictive analysis, not for text mining)	No (only suitable for numeric values, no text/structure mining)	No (just for numerical data)

Table 4.10: Commercial tools with trial option compared

	Maltego	Newprosoft Web Content Extractor	Data Applied
Developer/ Publisher	Paterva	Newprosoft	Data Applied
Homepage	`http://paterva.com/web6/`	`http://www.newprosoft.com/web-content-extractor.htm`	`http://www.data-applied.com`
Short description	Maltego is a crawling and mining tool that offers basically the following functions: - Find email addresses on a domain and see which resolve on social networks - Search for a specific email on the internet and see where it is used - Perform footprint on a domain (check for DNS names leading to this domain) - Search for twits on Twitter or status messages on Facebook - Try to find the geo location of a person on Twitter - Resolve affiliation of a Facebook or Twitter user including personal data like email, phone number, websites etc. - Get the network and domain information corresponding to an URL	This convenient to use crawler allows specifying crawling rules by selecting those elements on a website that should be extracted and those links that should be followed during the crawling process. This works especially well for websites that share the same structure.	Browser based mining tool that offers various filter and display techniques for numerical and discrete data like pivots, tree maps, correlation, displaying outliers, associations, clusters, decision trees or similarity maps. It can create predition models as well.
Import formats	web sites	web sites	Excel, csv, Salesforce.com, Dynamics CRM
Output formats	graph	.csv tables	Graph
Numerical data forecasting or model learning	No	No	Yes
(Past) numerical data pattern analysis	No	No	Yes
Runs stand-alone	Yes	Yes	Yes
Runs on Windows	Yes	Yes	Yes
Supports (Web) structure mining	Yes, but not the way required by this thesis	No	No
Supports text mining	No	No	No
Supports Web crawling	Yes but not freely definable as required by this thesis	Yes	No
Trial restrictions	result set limited to 12 (instead of 10.000 in commercial version)	14 days trial period, limited to 150 records	max. 500 rows, 1 workspace, 1 task at a time
Commercial version cost	Commercial 650$ first year, 320$ p.a. every following year	99 $	price unavailable ("buy" page malfunctual 02/13)
Qualifies for the analysis goal of this thesis	No (very mighty tool for gathering personal information, relationships and origins, especially for Facebook and Twitter but it lacks functionality required for this thesis)	Yes	No (handles only numerical values, no text or structure mining functionality)

Table 4.11: Commercial tools with trial option compared

	Rapid Miner	STATISTICA	XSight
Developer/ Publisher	Rapid-I	StatSoft	QSR International
Homepage	http://rapid-i.com/content/view/181/190/	http://www.statsoft.com/products/statistica-data-miner/	http://www.qsrinternational.com/products_xsight.aspx
Short description	Graph based workflow modelling tool. >500 operators for data transformation, dependency recognition, model creation, data classification, validation, visualization and more. Installable extensions provided by Rapid-I add further functionality for text analysis, PMML or WEKA support, enhanced reporting, text or web extension and more. Rapid-I offers the coupled software Rapid Analytics that allows storing data online and accessing many of the functions over the web browser.	This data mining tool offers comprehensive data mining methods (clustring, neural networks, classification, regression, SVM, association and sequence analysis, …). It offers wizard making offering results for less experienced users. With an optional extension it can be used for text mining and sentiment analysis as well.	XSight is a document analysis software that allows users to create different views on document data. It supports creating graphical maps that help organizing and categorizing text parts like in a flip chart, manual evaluation of text parts or commentaries, filtering and marking/querying text to help finding out specific information. The tool is optimized for surveys. Text mining is only possible in a fully manual way.
Import formats	Excel, Access, Oracle, SQL, Sybase, DB2, Ingrs, Postgres, SPSS, dBase, text files, Excel, SAS…	database entries over OLE*; Excel tables	documents (.doc, .rtf)
Output formats	graph, chart, diagram, table, textfile	database entries, PMML, models in C or Java, multiple graphical representations	documents (.doc, .ppt)
Numerical data forecasting or model learning	Yes	Yes	No
(Past) numerical data pattern analysis	Yes	Yes	No
Runs stand-alone	Yes	Yes	Yes
Runs on Windows	Yes	Yes	Yes, but only XP or Vista
Supports (Web) structre mining	No	Yes	No
Supports text mining	Yes (with extension)	Yes (with extension)	Yes, but only in a completely manual way
Supports Web crawling	Yes (with extension)	No	No
Trial restrictions	limited in service and extension functionality, no commercial use	30 days trial period; only basic functionality and just for prospective buyers, educational version is not free	30 days trial period
Commercial version cost	price unknown	depending on functionality 946 - 2.374 €	80 € per Semester, 150 € per year or 375 € unlimited for education; 1125 € for commercial use
Qualifies for the analysis goal of this thesis	Yes	No (no free evaluation version for students)	No (Text mining is only possible in a manual way that is supported by different text views)

Table 4.12: Commercial tools with trial option compared

	NVivo 10	Winweb Crawler v 2.0
Developer/ Publisher	QSR International	Newprosoft
Homepage	`http://www.qsrinternational.com/products_nvivo.aspx`	`http://www.winwebcrawler.com/index.htm`
Short description	Nvivo 10 is a document organizing and analyzing tool very similar to Xsight, but offers enhanced features. It can handle way more data formats and allows adding notes, highlighting and categorizing data parts. It can capture websites as PDF or data sets (table), select picture or video parts, comment them, do categorization and queries on selected data as well as creating reports just like in XSight.	An easy to use web crawler that is essentially able to extract URL, meta tags, plain text between <body> tags from Web sites, directories, search results or a list of URLs from a file.
Import formats	documents (.doc, .pdf, .rtf, .txt), tables (access, ODBC Excel), audio , video, pictures, webpages & web content	text file, search terms
Output formats	Graph, Tagcloud, other	.csv or text file
Numerical data forecasting or model learning	No	No
(Past) numerical data pattern analysis	No	No
Runs stand-alone	Yes	Yes
Runs on Windows	Yes	Yes
Supports (Web) structre mining	No	No
Supports text mining	Yes, but only in a completely manual way	No
Supports Web crawling	No	Yes
Trial restrictions	30 days trial period	15 days trial period
Commercial version cost	470 € (education) to 1645 (commercial)	99$
Qualifies for the analysis goal of this thesis	No (Text mining is only possible in a manual way that is supported by different text views)	Yes

Table 4.13: Commercial tools with trial option compared

4.3.4. Available Commercial Tools without Trial Option

Commercial tools disqualify from the start away for the practical testing phase because they do not offer any demo version and therefore cannot be installed without paying a fee that can be above 10.000 Euros making those tools a possible choice for larger companies only.

	Angoss Knowledge EXCELERATOR, SEEKER and STUDIO	**Coheris SPAD**
Developer/ Publisher	Angoss	Coheris
Homepage	`http://www.angoss.com/about-angoss/why-angoss`	`http://www.coheris.fr/en/page/produits/Spad.html`
Short description	These three software versions differentiate themselves in the number of features they offer. Knowledge EXCELERATOR is capable of data visualization and decision trees, SEEKER adds data preparation (joining, partitioning, etc.), prediction model validation and deployment as well as text mining and the STUDIO supports advanced modelling (neural networks, scorekards, market basket analysis, ...) as well as unsupervised model learning	This data mining tool from the french developer Coheris is designed to discover customer characteristics, most suitable communication channels for a company's target group, anticipate purchasing behaviour, improve effectiveness of customer retention, fraud prevention and production process improvement. It is able to perform text mining as well in form of term frequency and term distance measures, etc. Term distance text mining can be used e.g. for email classification.
Import formats	database entries, tables, plain text (with text analytics addon)	Database entries (Oracle, Sybase, SAS, SQL Server, DB2), ODBC, text files, Excel, SPSS***
Output formats	table, graph, SQL, XML, PMML**, SAS	graph, diagram, Excel table
Numerical data forecasting or model learning	Yes	Yes
(Past) numerical data pattern analysis	Yes	Yes
Runs stand-alone	Yes	Yes
Runs on Windows	Yes	Yes
Supports (Web) structure mining	Yes	No
Supports text mining	Yes	Yes
Supports Web crawling	No	No
Commercial version cost	Angoss Support did not respond on query, they just stated that there is no free trial version in November 2012.	unknown (customer area just in French; it is possible to demand web demo there but the form required a French telephone number)
Would qualify for analysis goal if software was free	Yes	Yes

Table 4.14: Pure commercial data mining tools compared

	RapidSentilyzer	RapidNet
Developer/ Publisher	Rapid-I	Rapid-I
Homepage	`http://rapid-i.com/content/view/184/194/lang,en/`	`https://rapid-i.com/content/view/183/193/lang,de/`
Short description	A software based on the RapidMiner engine that automatically collects the latest news about a company, its products as well as its competitors and provides different views on the data through a browser-based app. It allows sentiment predictions for single texts, sntiment statistics for market segment as well as insight in the reasons for specific sentiments. It is configured and maintained by Rapid-I staff.	Based on RapidMiner this software is able to explore structures. It is targeted for structure and relationship finding between a company and its customers, deliveries and their destinations, business process components, etc. It allows to display geographic information on maps - similar to tableau public. Free version cannot be downloaded from Rapid-I, you have to make an inquiry.
Import formats	detailed company data is forwarded to Rapid-I staff in a human readable form	Excel, Access, Oracle, SQL, Sybase, DB2, Ingres, Postgres, SPSS, dBase, text files, Excel, SAS…
Output formats	graph, chart, diagram	graph, geographical map
Numerical data forecasting or model learning	No	No
(Past) numerical data pattern analysis	No	No
Runs stand-alone	Yes	Yes
Runs on Windows	Yes	Yes
Supports (Web) structure mining	No	Yes
Supports text mining	Yes	No
Supports Web crawling	Yes	No
Commercial version cost	unknown (depends on company requirements and project size - data mining job is done by company staff)	unknown
Would qualify for analysis goal if software was free	Yes	Yes

Table 4.15: Pure commercial data mining tools compared

	SAS	SPSS Modeler	Oracle Data Mining
Developer/ Publisher	SAS Institute Inc.	IBM Corp.	Oracle
Homepage	`http://www.sas.com/technologies/analytics/datamining/miner/`	`http://www-142.ibm.com/software/products/at/de/spss-modeler/`	`http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/index.html`
Short description	A very powerful yet expensive data mining tool usings its own proprietary data format but supports third party databases as well. It focusses on predictive and descriptive models for market basket analysis, decision trees, regression, neural networks, time series analys, incremental response models and more.	SPSS Modeler is designed for companies to get an understanding for their customer relationships. It has an integrated text analsyis workbench that can perform analysis tasks on text from documents and online sources. It can extract entities and sentiments and has a buildt-in translator for many languages including Chinese.	Oracle Data Mining is a data mining extension for ORACLE databases that provides algorithms for classification, prediction, regression, association mining, feature selection, anomality detection and feature extraction. It needs ORACLE database in order to run.
Import formats	SAS, netezza[23], db2, aster ncluster[24], Teradata[25], ORACLE, Access, Excel, .csv	ODBC databases, .csv, SAS, SPSS	ORACLE
Output formats	SAS, C, Java, PMML, Excel, Acess, Graphs	SPSS, Excel, Graph, SPSS	Graph, ORACLE
Numerical data forecasting or model learning	Yes	Yes	Yes
(Past) numerical data pattern analysis	Yes	Yes	Yes
Runs stand-alone	Yes	Yes	Yes but it needs ORACLE DB
Runs on Windows	Yes	Yes	Yes
Supports (Web) structure mining	Yes	Yes	No
Supports text mining	Yes	Yes	Yes
Supports Web crawling	Yes	Yes	No
Commercial version cost	approximately 20.000 $	10.000 - 120.000 € depending on user count and version + 12 month support	With ORACLE Enterprise Edition 1.200 - 2.000 € per month.
Would qualify for analysis goal of this thesis if software was free	Yes	Yes	Yes

Table 4.16: Pure commercial data mining tools compared

[23] Analytic database and data warehouse application by a subsidiary of IBM
[24] Analytic database for frontline data warehousing owned by TERADATA
[25] Relational datbase management system (RDBMS) developed by Teradata and owned by NCR Corporation

4.4. Selected Crawling Tools

Out of all 13 free tools, 11 commercial tools with trial option and 7 pure commercial tools under review 4 free, 4 trial and 3 commercial tools offer web crawling abilities. 3 trial tools managed to qualify for the practical test phase: Newprosoft Web Content Extractor, RapidMiner and Winweb Crawler v 2.0. While RapidMiner offers sophisticated web mining functionality too, both other tools are decided web crawlers. The trial version of Web Content Extractor is limited in the number of analysable data sets (maximal 150), while RapidMiner and WinwebCrawler have only time and service constraints that do not affect functionality.

Note that all 3 commercial crawler would have qualified as well if they had offered some free test version.

4.5. Selected Mining Tools

From all 13 free tools, 11 commercial tools with trial option and 7 pure commercial tools under reviews the majority of 10 free, 9 trial and all 7 commercial tools offered data mining functionality. However, only 2 free tools and 1 trial tool could manage to qualify for practical testing: KNIME, Orange and RapidMiner. RapidMiner has crawling functionality too while KNIME and Orange offer data mining functionality only. Since KNIME and Orange are free to use they can be used without any restrictions.

All 7 commercial miners would have qualified for practical testing if some free test versions were available. Investing money in this project would have doubled the number of suitable crawling software and more than tripled the number of suitable mining software candidates.

5. Crawling Tool Tests

In this chapter the data collection process on communities will be described. Figure 5.1 shows an overview how the three selected crawling tools are used on both communities to get the best result datasets.

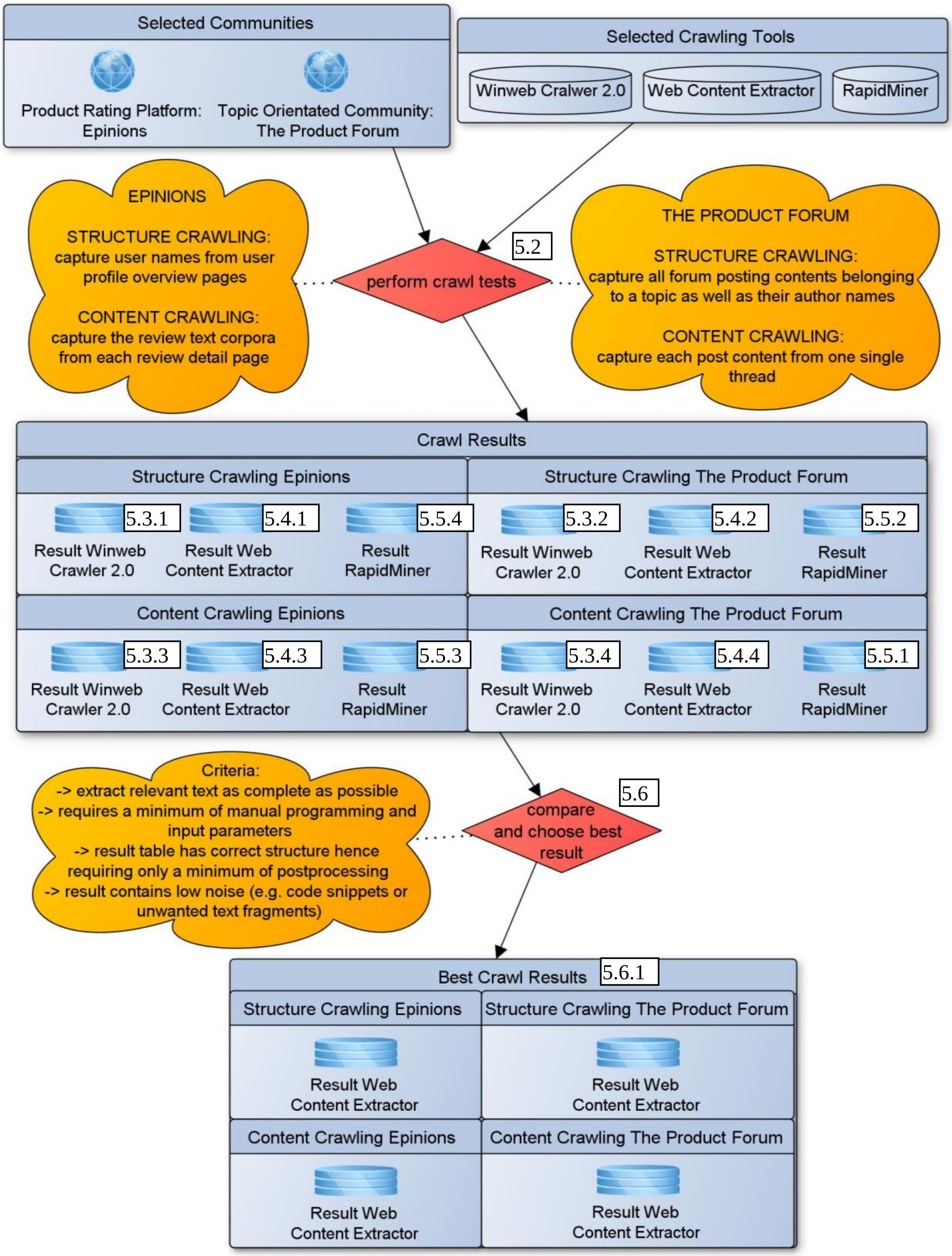

Selected Communities
Product Rating Platform: Epinions
Topic Orientated Community: The Product Forum
Selected Crawling Tools
Winweb Cralwer 2.0
Web Content Extractor
RapidMiner
EPINIONS
STRUCTURE CRAWLING: capture user names from user profile overview pages
CONTENT CRAWLING: capture the review text corpora from each review detail page
perform crawl tests
5.2
THE PRODUCT FORUM
STRUCTURE CRAWLING: capture all forum posting contents belonging to a topic as well as their author names
CONTENT CRAWLING: capture each post content from one single thread
Crawl Results
Structure Crawling Epinions
5.3.1 Result Winweb Crawler 2.0
5.4.1 Result Web Content Extractor
5.5.4 Result RapidMiner
Structure Crawling The Product Forum
5.3.2 Result Winweb Crawler 2.0
5.4.2 Result Web Content Extractor
5.5.2 Result RapidMiner
Content Crawling Epinions
5.3.3 Result Winweb Crawler 2.0
5.4.3 Result Web Content Extractor
5.5.3 Result RapidMiner
Content Crawling The Product Forum
5.3.4 Result Winweb Crawler 2.0
5.4.4 Result Web Content Extractor
5.5.1 Result RapidMiner
Criteria:
-> extract relevant text as complete as possible
-> requires a minimum of manual programming and input parameters
-> result table has correct structure hence requiring only a minimum of postprocessing
-> result contains low noise (e.g. code snippets or unwanted text fragments)
compare and choose best result
5.6
Best Crawl Results
5.6.1
Structure Crawling Epinions
Result Web Content Extractor
Structure Crawling The Product Forum
Result Web Content Extractor
Content Crawling Epinions
Result Web Content Extractor
Content Crawling The Product Forum
Result Web Content Extractor

Figure 5.1: Overview of the crawling process performed in this chapter. Each number in a white rectangle describes the chapters where the corresponding task or result is described in further detail.

5.1. Test Configuration

The test configuration was as shown in table 5.1 and did not change during the whole test procedure.

Processor	i5-3570, 4x3,4GHz
RAM	8 GB DDR3 1600MHz
Mainboard	Gigabyte GA-B75N, Socket 1155
Harddrive	256 GB Crucial M4 SSD
Operating System	Windows 7 Ultimate, 64 Bit

Table 5.1: Test hardware used for installing and using all tools

For mobile working and in order to reproduce specific software behaviour on a second system, the platform shown below was used.

Processor:	i5-2410M 2x2,6 GHz
RAM:	4 GB DDR3 1333MHz
Mainboard	Proprietary Sony Vaio S, Socket 1155
Harddrive	120 GB OCZ Vertex 3 SSD
Operating System	Windows 7 Home Premium, 64 Bit

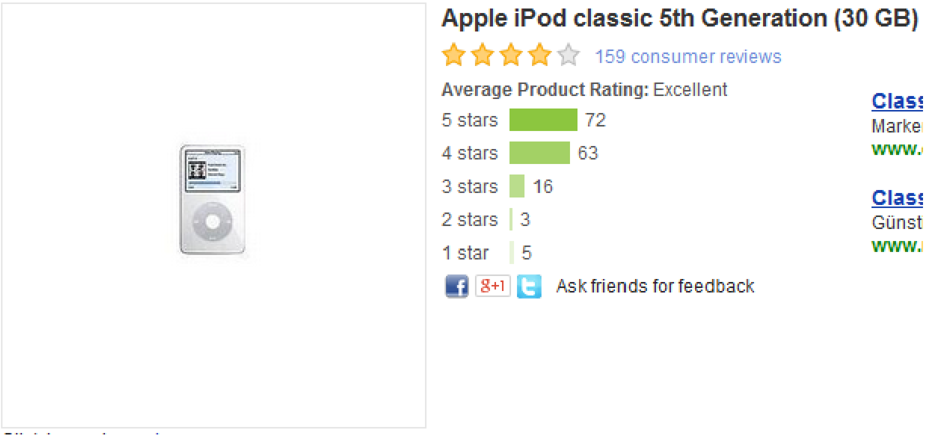
Table 5.2: Second test system used for some subtasks and to test software behaviour on a second system

5.2. General Crawling Process and desired Results

Despite any tool has its own features, different search algorithms, user interfaces and result formats, the basic crawling process works very similar with each of them. Each crawler has to perform four crawling tasks as shown in figure 5.1. The following two subchapters will describe the crawling process and describe what the desired results should ideally look like. At the End of chapter 5, after all tools have been tested, results will be evaluated.

5.2.1. Structure Crawling Epinions

The overall review score distribution for the device under review is shown in figure 5.2.



Apple iPod classic 5th Generation (30 GB)

★★★★☆ 159 consumer reviews

Average Product Rating: Excellent

5 stars 72

4 stars 63

3 stars 16

2 stars 3

1 star 5

Ask friends for feedback

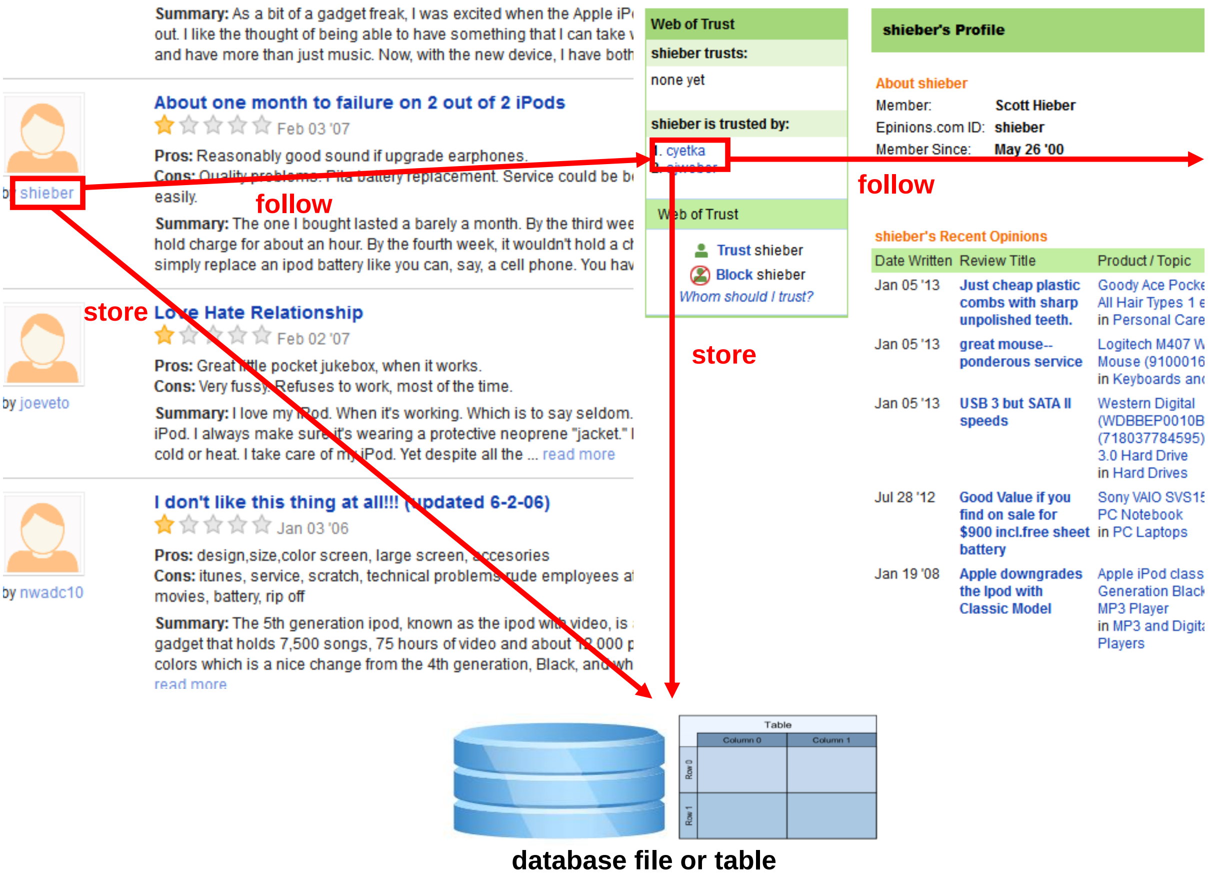
Class
Marke
www.

Class
Günst
www.

Figure 5.2: Rating distribution overview for iPod clasic shown on the Epinions website.

Structure crawling Epinions was done for all three users that posted negative full text reviews. Negative reviews are those where the reviewer gave only one star on the available scale from 1 to 5. The 2 reviews missing to the number of 5 shown in the figure have not written a long review linked with a user profile and therefore cannot be evaluated. Negative reviews are chosen since the overall sentiment across the whole 159 reviews is very positive raising the question if the authors of those three negative reviews share some special attributes influencing their trustworthiness. Those authors are the three users named shieber, joveto and nwadc10. Looking at them shieber is the most active user. He has written 35 reviews and is trusted by two users. The question to be answered is: "Are those two users trustworthy?" This will be discovered by analysing the two trustees as well as those users expressing trust to them.

The default starting URL for crawlers is the product rating overview page containing a complete list of all review headings with abstracts sorted by rating points.[26] By clicking on the names a detailed user profile opens showing who trusts the actual user and which users are trusted by him. The goal of the structure crawling task on Epinions is to extract those names and return a list of users and their trust-relationships in a form "user A trusts user B" as described and shown in chapter 4.2.1 and figure 4.1. Alternatively the crawl can be started directly from the three user profile sites of shieber, joveto and nwadc10.



Summary: As a bit of a gadget freak, I was excited when the Apple iP
out. I like the thought of being able to have something that I can take
and have more than just music. Now, with the new device, I have both

by shieber

About one month to failure on 2 out of 2 iPods
Feb 03 '07

Pros: Reasonably good sound if upgrade earphones.
Cons: Quality problems. Pita battery replacement. Service could be b
easily.

Summary: The one I bought lasted a barely a month. By the third wee
hold charge for about an hour. By the fourth week, it wouldn't hold a c
simply replace an ipod battery like you can, say, a cell phone. You hav

by joeveto

Love Hate Relationship
Feb 02 '07

Pros: Great little pocket jukebox, when it works.
Cons: Very fussy. Refuses to work, most of the time.

Summary: I love my iPod. When it's working. Which is to say seldom.
iPod. I always make sure it's wearing a protective neoprene "jacket." I
cold or heat. I take care of my iPod. Yet despite all the ... read more

by nwadc10

I don't like this thing at all!!! (updated 6-2-06)
Jan 03 '06

Pros: design,size,color screen, large screen, accesories
Cons: itunes, service, scratch, technical problems rude employees a
movies, battery, rip off

Summary: The 5th generation ipod, known as the ipod with video, is
gadget that holds 7,500 songs, 75 hours of video and about 12,000 p
colors which is a nice change from the 4th generation, Black, and wh
read more

Web of Trust

shieber trusts:
none yet

shieber is trusted by:
1. cyetka

Web of Trust
Trust shieber
Block shieber
Whom should I trust?

shieber's Profile

About shieber

Member:	**Scott Hieber**
Epinions.com ID:	**shieber**
Member Since:	**May 26 '00**

shieber's Recent Opinions

Date Written	Review Title	Product / Topic
Jan 05 '13	**Just cheap plastic combs with sharp unpolished teeth.**	Goody Ace Pocke All Hair Types 1 e in Personal Care
Jan 05 '13	**great mouse-- ponderous service**	Logitech M407 W Mouse (9100016 in Keyboards and
Jan 05 '13	**USB 3 but SATA II speeds**	Western Digital (WDBBEP0010B (718037784595) 3.0 Hard Drive in Hard Drives
Jul 28 '12	**Good Value if you find on sale for $900 incl.free sheet battery**	Sony VAIO SVS15 PC Notebook in PC Laptops
Jan 19 '08	**Apple downgrades the Ipod with Classic Model**	Apple iPod class Generation Black MP3 Player in MP3 and Digit Players

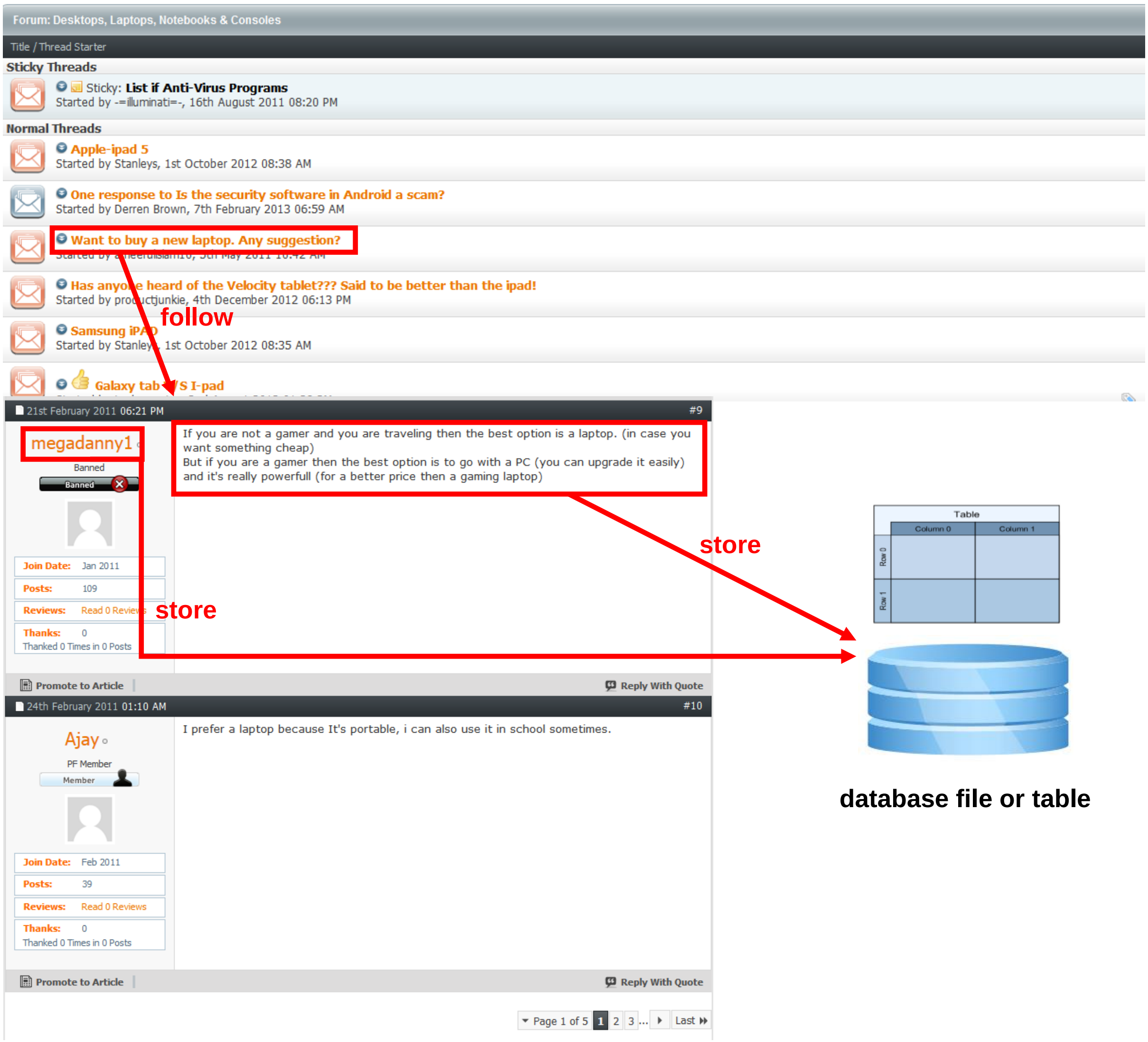
Figure 5.3: Reviews on Epinions as shown on the overview page forming the crawler starting point (left). The user profile detail page that can be called by clicking user names on the overview page is shown on the right. By following the "cyetka" link in this example "cyetka`s profile is opened.

[26] Starting URL: `http://www.epinions.com/reviews/Apple_iPod_White_30_GB_MA002LL_A_MP3_Player/sec_~opinion_list/pp_~1/pa_~1/sort_~prdrt/sort_dir_~des#list`

5.2.2. Structure Crawling The Product Forum

Structure crawling The Product Forum works a bit different than on Epinions since not so much hyperlinks have to be followed but instead names from the posts have to be extracted. The desired result is a list of available user names and the corresponding post messages the user has written. Those user names mentioned in a post will be regarded as users this message is addressed to. The final “user A talks to user B” data structure will later be derived from that raw crawling results.

The crawler has to extract author names and their post content from every thread in the category “Desktops, Laptops, Notebooks and Consoles” shown on the overview-page.[27] It should ideally follow each thread link automatically collecting user-post combinations across the whole thread batch in that category. Note that this source page is constantly changing probably making results obtained in the later test part not exactly reproducible.



Forum: Desktops, Laptops, Notebooks & Consoles

Title / Thread Starter

Sticky Threads

Sticky: **List if Anti-Virus Programs**
Started by -=illuminati=-, 16th August 2011 08:20 PM

Normal Threads

Apple-ipad 5
Started by Stanleys, 1st October 2012 08:38 AM

One response to Is the security software in Android a scam?
Started by Derren Brown, 7th February 2013 06:59 AM

Want to buy a new laptop. Any suggestion?

Has anyo e heard of the Velocity tablet??? Said to be better than the ipad!
Started by pro uctjunkie, 4th December 2012 06:13 PM

Samsung iPA
Started by Stanley , 1st October 2012 08:35 AM

Galaxy tab /S I-pad

21st February 2011 06:21 PM #9

megadanny1
Banned

Join Date: Jan 2011
Posts: 109
Reviews: Read 0 Review
Thanks: 0
Thanked 0 Times in 0 Posts

If you are not a gamer and you are traveling then the best option is a laptop. (in case you want something cheap)
But if you are a gamer then the best option is to go with a PC (you can upgrade it easily) and it's really powerfull (for a better price then a gaming laptop)

24th February 2011 01:10 AM #10

Ajay
PF Member

Join Date: Feb 2011
Posts: 39
Reviews: Read 0 Reviews
Thanks: 0
Thanked 0 Times in 0 Posts

I prefer a laptop because It's portable, i can also use it in school sometimes.

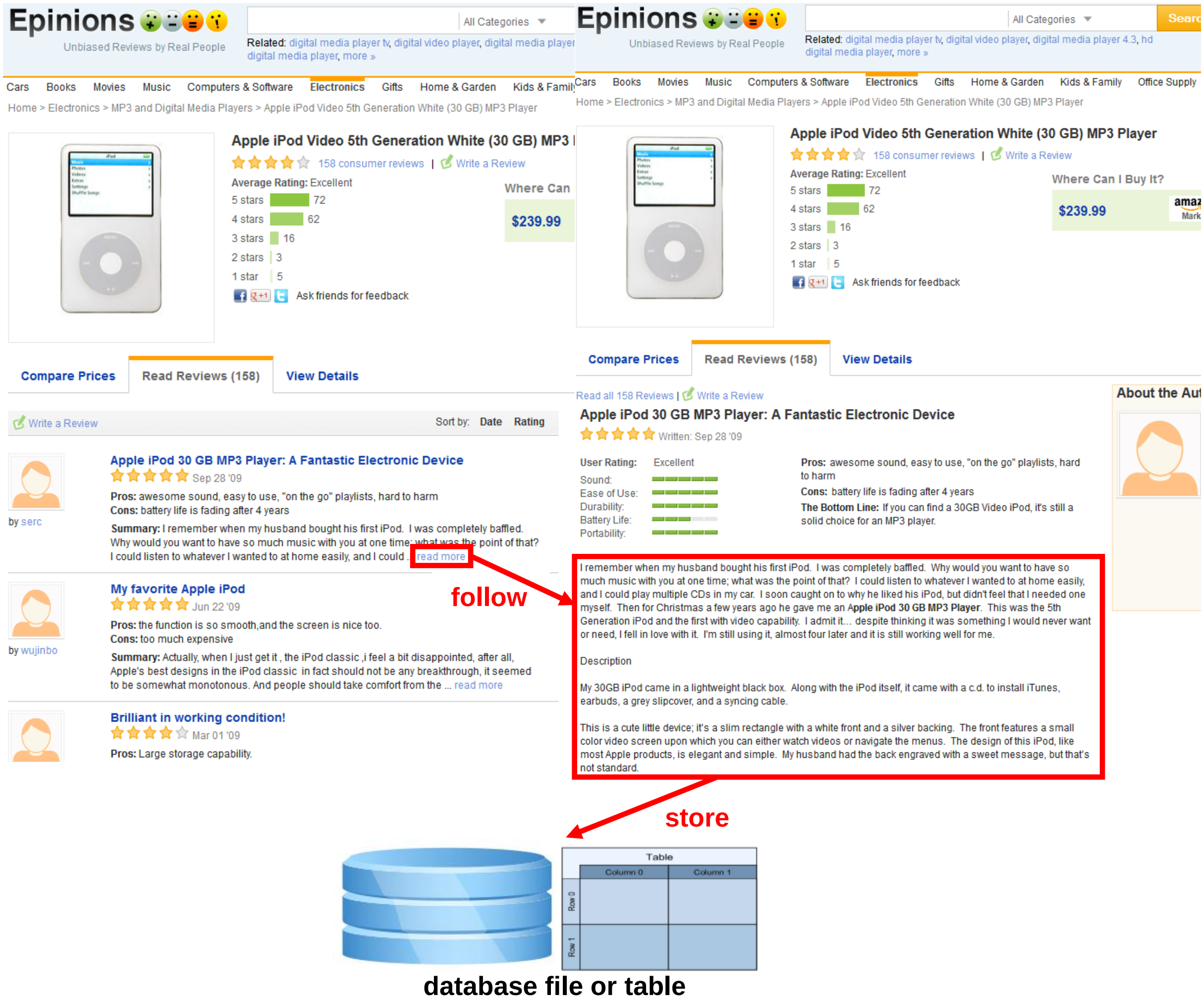
Figure 5.4: Top: thread overview page on The Product Forum – from here the structure crawling process starts. Bottom: from each single posts the author name and content is stored in a database or table.

[27] http://www.theproductforum.com/forums/f7/

5.2.3. Content Crawling Epinions

Content crawling Epinions starts from the review overview page just like the structure mining process did, except that this time there is no need to sort the results.[28] Following each “read more” link leads to the full review text that should be extracted by the crawler. Since short reviews lack such a link, each headline was selected that links to the full content as well and works for short reviews.

The whole sum of all detail reviews consisting mainly of plain text but also of specialised fields for positive or negative features and a summarizing sentence written in the field: “the bottom line” form a solid base for later analysis of features and in which way they are mentioned (positive or negative).



Epinions

Unbiased Reviews by Real People

All Categories

Related: digital media player tv, digital video player, digital media player digital media player, more »

Cars Books Movies Music Computers & Software Electronics Gifts Home & Garden Kids & Famil

Home > Electronics > MP3 and Digital Media Players > Apple iPod Video 5th Generation White (30 GB) MP3 Player

Apple iPod Video 5th Generation White (30 GB) MP3 I

★★★★☆ 158 consumer reviews | Write a Review

Average Rating: Excellent

5 stars	72
4 stars	62
3 stars	16
2 stars	3
1 star	5

Ask friends for feedback

Where Can

$239.99

Compare Prices | Read Reviews (158) | View Details

Write a Review Sort by: Date Rating

Apple iPod 30 GB MP3 Player: A Fantastic Electronic Device
★★★★★ Sep 28 '09
by serc
Pros: awesome sound, easy to use, "on the go" playlists, hard to harm
Cons: battery life is fading after 4 years
Summary: I remember when my husband bought his first iPod. I was completely baffled. Why would you want to have so much music with you at one time; what was the point of that? I could listen to whatever I wanted to at home easily, and I could ... read more

My favorite Apple iPod
★★★★★ Jun 22 '09
by wujinbo
Pros: the function is so smooth,and the screen is nice too.
Cons: too much expensive
Summary: Actually, when I just get it , the iPod classic ,i feel a bit disappointed, after all, Apple's best designs in the iPod classic in fact should not be any breakthrough, it seemed to be somewhat monotonous. And people should take comfort from the ... read more

Brilliant in working condition!
★★★★☆ Mar 01 '09
Pros: Large storage capability.

Epinions

Unbiased Reviews by Real People

All Categories Searc

Related: digital media player tv, digital video player, digital media player 4.3, hd digital media player, more »

Cars Books Movies Music Computers & Software Electronics Gifts Home & Garden Kids & Family Office Supply

Home > Electronics > MP3 and Digital Media Players > Apple iPod Video 5th Generation White (30 GB) MP3 Player

Apple iPod Video 5th Generation White (30 GB) MP3 Player

★★★★☆ 158 consumer reviews | Write a Review

Average Rating: Excellent

5 stars	72
4 stars	62
3 stars	16
2 stars	3
1 star	5

Ask friends for feedback

Where Can I Buy It?

$239.99 amaz Mark

Compare Prices | Read Reviews (158) | View Details

Read all 158 Reviews | Write a Review

Apple iPod 30 GB MP3 Player: A Fantastic Electronic Device
★★★★★ Written: Sep 28 '09

User Rating: Excellent
Sound:
Ease of Use:
Durability:
Battery Life:
Portability:

Pros: awesome sound, easy to use, "on the go" playlists, hard to harm
Cons: battery life is fading after 4 years
The Bottom Line: If you can find a 30GB Video iPod, it's still a solid choice for an MP3 player.

About the Aut

I remember when my husband bought his first iPod. I was completely baffled. Why would you want to have so much music with you at one time; what was the point of that? I could listen to whatever I wanted to at home easily, and I could play multiple CDs in my car. I soon caught on to why he liked his iPod, but didn't feel that I needed one myself. Then for Christmas a few years ago he gave me an **Apple iPod 30 GB MP3 Player**. This was the 5th Generation iPod and the first with video capability. I admit it... despite thinking it was something I would never want or need, I fell in love with it. I'm still using it, almost four later and it is still working well for me.

Description

My 30GB iPod came in a lightweight black box. Along with the iPod itself, it came with a c.d. to install iTunes, earbuds, a grey slipcover, and a syncing cable.

This is a cute little device; it's a slim rectangle with a white front and a silver backing. The front features a small color video screen upon which you can either watch videos or navigate the menus. The design of this iPod, like most Apple products, is elegant and simple. My husband had the back engraved with a sweet message, but that's not standard.

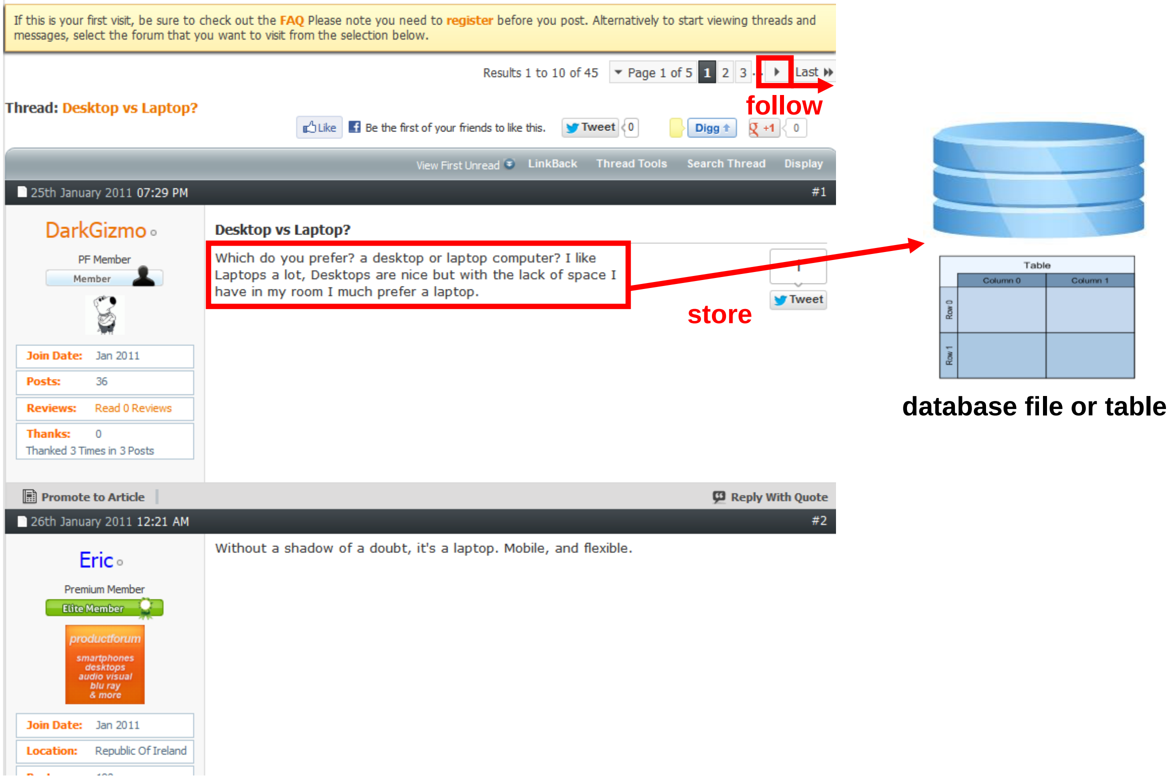
Figure 5.5: The top left screen shows the product overview page while the top right one displays the detailed review that is shown after following the “read more” link. Each text corpora on this site has to be extracted and stored in a database or table.

[28] `http://www.epinions.com/reviews/Apple_iPod_White_30_GB_MA002LL_A_MP3_Player/sec_~opinion_list/pp_~1/pa_~1?sb=1`

5.2.4. Content Crawling The Product Forum

Content crawling The Product Forum is fairly easy as just the content from one single thread has to be extracted. The only difficulty occurs from the fact that the forum page only displays 10 postings per page. Since the thread under review “Desktop vs. Laptop” has well over 40 postings it is necessary to follow those “next” links situated on the top or bottom of the website as can be seen in Figure 5.6.

By default crawling starts from the first thread page.[29] The desired crawl result is plain text from the body of the post messages while opposed to the structure mining task post author names are not needed and therefore left out.



If this is your first visit, be sure to check out the FAQ Please note you need to register before you post. Alternatively to start viewing threads and messages, select the forum that you want to visit from the selection below.

Results 1 to 10 of 45 Page 1 of 5 1 2 3 ... Last

Thread: Desktop vs Laptop?

25th January 2011 07:29 PM #1

DarkGizmo
PF Member
Member

Join Date: Jan 2011
Posts: 36
Reviews: Read 0 Reviews
Thanks: 0
Thanked 3 Times in 3 Posts

Desktop vs Laptop?

Which do you prefer? a desktop or laptop computer? I like Laptops a lot, Desktops are nice but with the lack of space I have in my room I much prefer a laptop.

26th January 2011 12:21 AM #2

Eric
Premium Member
Elite Member

Join Date: Jan 2011
Location: Republic Of Ireland

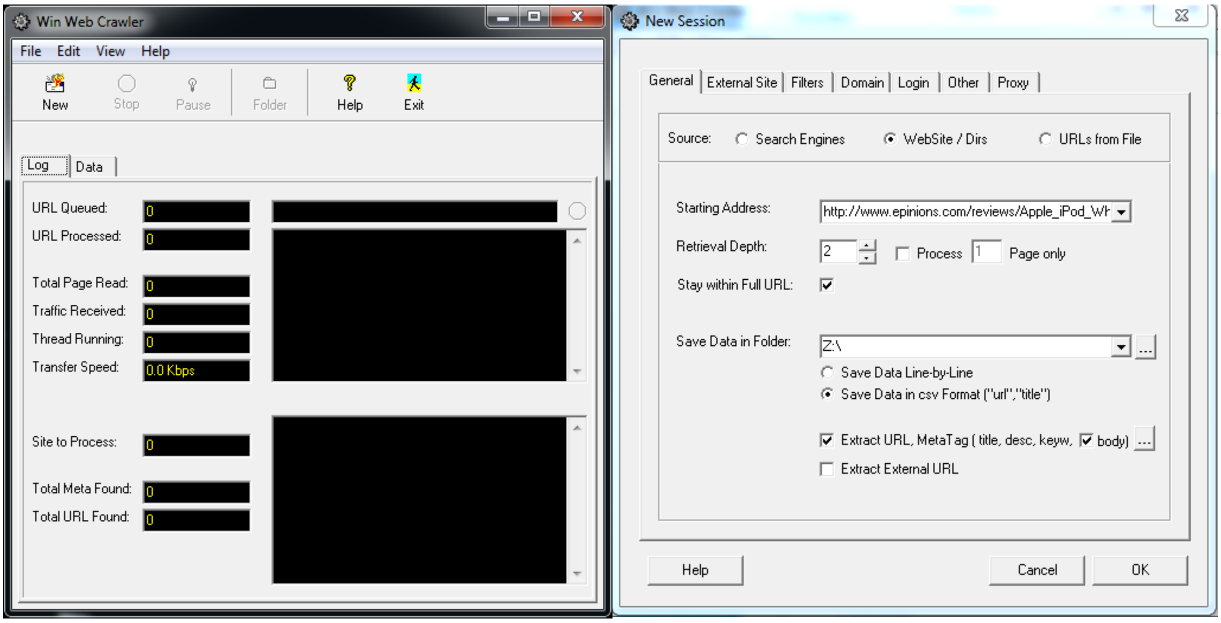
Without a shadow of a doubt, it's a laptop. Mobile, and flexible.

Figure 5.6: The upper end of the The Product Forum thread page showing “next page” button the crawler has to follow. All post contents are extracted and stored in database or table files.

[29] `http://www.theproductforum.com/forums/f7/desktop-vs-laptop-214.html`

5.3. Winweb Crawler v2.0

The **installation of** Winweb Crawler is very easy: just download the installer file from `http://www.winwebcrawler.com/download.htm`, execute the file and you are done.


Win Web Crawler
File Edit View Help
New Stop Pause Folder Help Exit
Log Data
New Session
General External Site Filters Domain Login Other Proxy
Help Cancel OK

URL Queued: 0
URL Processed: 0
Total Page Read: 0
Traffic Received: 0
Thread Running: 0
Transfer Speed: 0.0 Kbps
Site to Process: 0
Total Meta Found: 0
Total URL Found: 0

Source: ○ Search Engines ● WebSite / Dirs ○ URLs from File
Starting Address: http://www.epinions.com/reviews/Apple_iPod_Wh
Retrieval Depth: 2 □ Process 1 Page only
Stay within Full URL: ☑
Save Data in Folder: Z:\
○ Save Data Line-by-Line
● Save Data in csv Format ("url","title")
☑ Extract URL, MetaTag (title, desc, keyw, ☑ body)
□ Extract External URL

Figure 5.7: User interface of Newprosoft Win Web Crawler

Function principle: A crawling task is initiated by either specifying a starting URL from which the program will collect all data and follow hyperlinks that lead to further URLs up to a specified search retrieval depth. E.g. a retrieval depth of 2 means every hyperlink on the starting site A is followed as well as every hyperlink on those link targets (e.g. page A links to B and C that in turn link to D, E and F) but none further. To refine the search process it can be specified whether URLs or pages must or must not contain specific words or text parts in order to be crawled. Alternatively a text file containing a list of given URLs can be used. The results are stored in a .csv table.

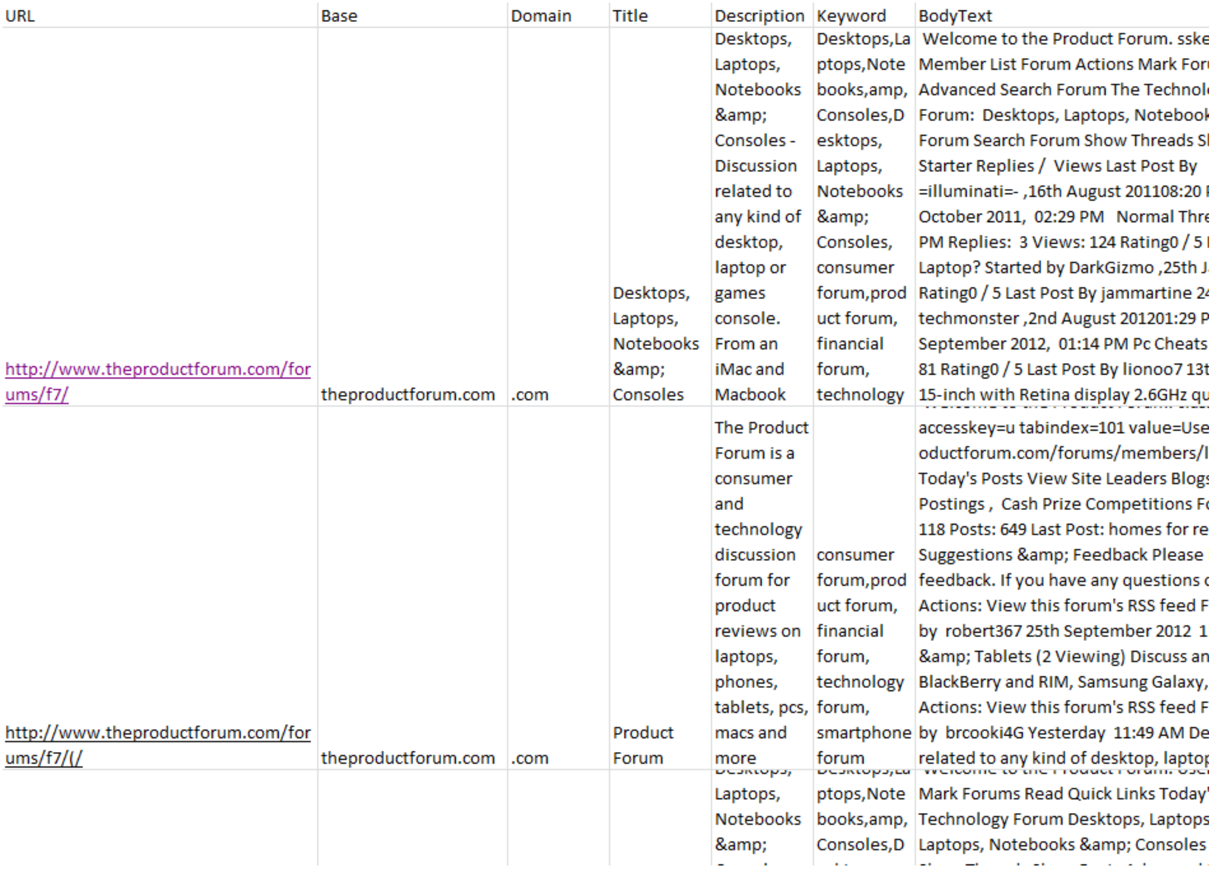
5.3.1. Structure Crawling Epinions

Despite any tests run while using the overview list, a profile (e.g. the user profile detail page of “shieber”) or even an URL list containing all user profiles that are of interest for this analysis, no result could be obtained since the program only returned an empty file regardless which other parameters where specified.

5.3.2. Structure Crawling The Product Forum

Starting from the thread overview page, usable results could be obtained through setting “stay within full URL” meaning only pages containing the full text from the starting URL were considered. A retrieval depth of 5 delivered the best result completeness to noise ratio. URLs with “?” and “print” were excluded to get rid of PHP-variable modified and printable website views and URLs were forced to include .html which excluded all thread overview sites.

The result was usable but had some flaws. Some sub-pages of threads were left out – e.g. in the “Desktops vs. Laptops” thread only 1 of 5 pages were extracted. On the other hand all surrounding text like menu entries or commercials and tags were present in the crawling result making it noisy.



URL	Base	Domain	Title	Description	Keyword	BodyText
http://www.theproductforum.com/forums/f7/	theproductforum.com	.com	Desktops, Laptops, Notebooks &amp; Consoles	Desktops, Laptops, Notebooks &amp; Consoles - Discussion related to any kind of desktop, laptop or games console. From an iMac and Macbook	Desktops,Laptops,Notebooks,amp,Consoles,Desktops, Laptops, Notebooks &amp; Consoles, consumer forum,product forum, financial forum, technology	Welcome to the Product Forum. sske Member List Forum Actions Mark For Advanced Search Forum The Technol Forum: Desktops, Laptops, Notebook Forum Search Forum Show Threads S Starter Replies / Views Last Post By =illuminati=- ,16th August 201108:20 October 2011, 02:29 PM Normal Thre PM Replies: 3 Views: 124 Rating0 / 5 Laptop? Started by DarkGizmo ,25th J Rating0 / 5 Last Post By jammartine 2 techmonster ,2nd August 201201:29 P September 2012, 01:14 PM Pc Cheats 81 Rating0 / 5 Last Post By lionoo7 13t 15-inch with Retina display 2.6GHz qu
http://www.theproductforum.com/forums/f7/(/	theproductforum.com	.com	Product Forum	The Product Forum is a consumer and technology discussion forum for product reviews on laptops, phones, tablets, pcs, macs and more	consumer forum,product forum, financial forum, technology forum, smartphone forum	accesskey=u tabindex=101 value=Use oductforum.com/forums/members/l Today's Posts View Site Leaders Blog Postings , Cash Prize Competitions F 118 Posts: 649 Last Post: homes for re Suggestions &amp; Feedback Please feedback. If you have any questions Actions: View this forum's RSS feed F by robert367 25th September 2012 1 &amp; Tablets (2 Viewing) Discuss an BlackBerry and RIM, Samsung Galaxy, Actions: View this forum's RSS feed F by brcooki4G Yesterday 11:49 AM De related to any kind of desktop, lapto
				Desktops, Laptops, Notebooks &amp;	Desktops,Laptops,Notebooks,amp,Consoles,D	Welcome to the Product Forum. Use Mark Forums Read Quick Links Today' Technology Forum Desktops, Laptops Laptops, Notebooks &amp; Consoles

Figure 5.8: Structure crawl results from The Product Forum delivered by Winweb Crawler 2.0

5.3.3. Content Crawling Epinions

Starting from the slightly modified default starting web address[30] several test runs were performed with varying retrieval depth between 1 and 30. The option “stay within the full URL” was always on.

The crawler was able to extract the given product overview site quite completely. Unfortunately, despite various attempts to do so, there was no way to automatically extract the complete summary texts since the crawler refused to follow the corresponding links but rather landed on the same overview sites for several hundred times and extracted the same body text over and over again. In other tests with similar webpages it followed the links correctly so this problem seems to occur only on this specific webpage.

With a manual created URL list the content could be extracted but this means calling every review by hand and copying the link into a text file. This is nearly as time consuming as copying every review in a table by hand making it possible for small projects but impractical.

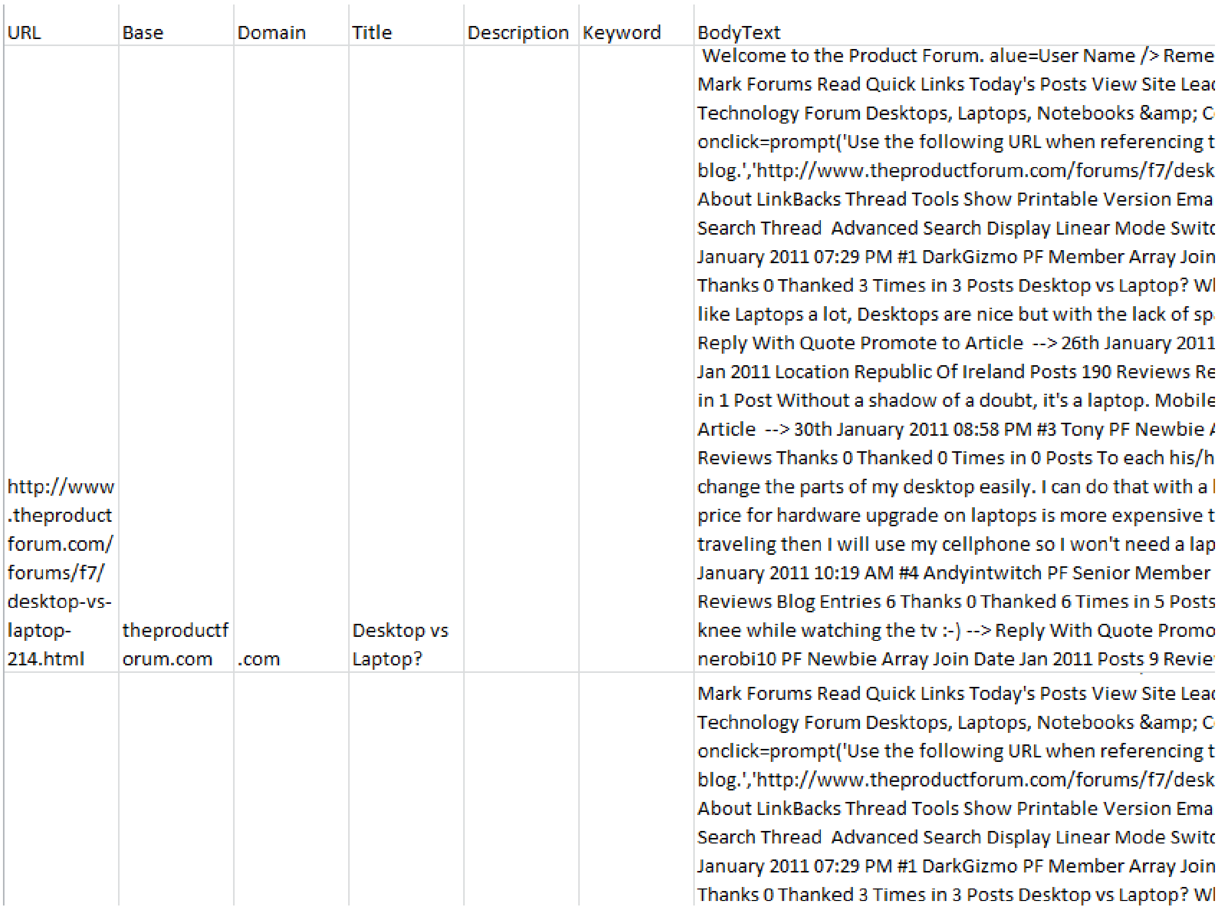
Regarding the result quality, since no data cleaning or pre-selection of the crawled data can be done with the program, every single tag and text fragment between the two `<body></body>` tags of the website was extracted leading to a rather overwhelming and unstructured amount of text that begs for cleaning and postprocessing.

[30] `http://www.epinions.com/reviews/Apple_iPod_White_30_GB_MA002LL_A_MP3_Player`

5.3.4. Content Crawling The Product Forum

Starting from the default starting URL several test runs were conducted using different settings. The selected retrieval depth had no effect at all just like the option “stay within full URL”. To get cleaner results, PHP modified URLs including “?” and print versions of sites including “print” were excluded with the build-in filter.

The results were acceptable since everything relevant was extracted from the 5 pages under review (containing each ~10 thread posts). Unfortunately way too much was extracted: duplicate data sets and many text fragments that would not belong to post messages. Data cleaning and post processing has to be performed to make data usable for further data mining.

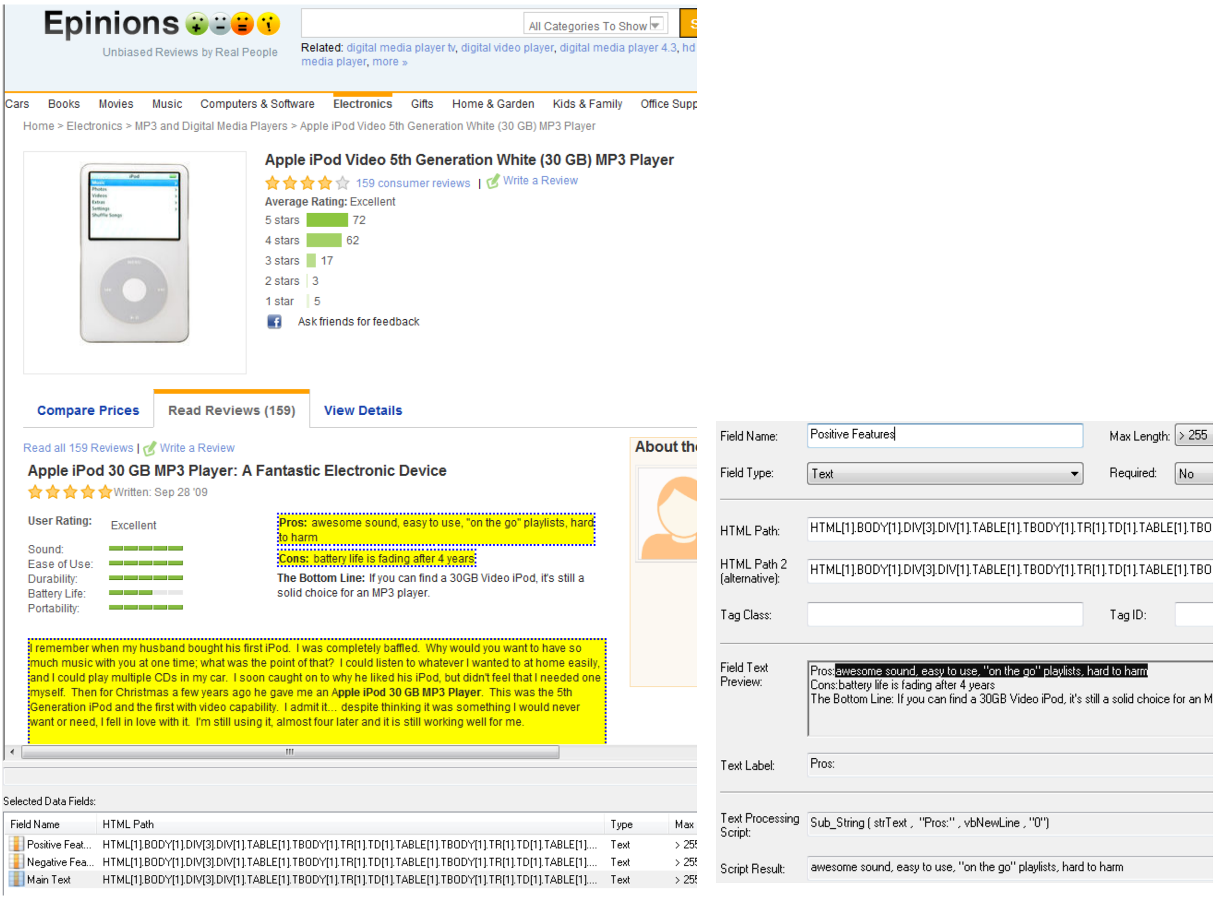


URL	Base	Domain	Title	Description	Keyword	BodyText
http://www.theproductforum.com/forums/f7/desktop-vs-laptop-214.html	theproductforum.com	.com	Desktop vs Laptop?			Welcome to the Product Forum. alue=User Name /> Reme Mark Forums Read Quick Links Today's Posts View Site Lea Technology Forum Desktops, Laptops, Notebooks &amp; C onclick=prompt('Use the following URL when referencing t blog.','http://www.theproductforum.com/forums/f7/desk About LinkBacks Thread Tools Show Printable Version Ema Search Thread Advanced Search Display Linear Mode Switc January 2011 07:29 PM #1 DarkGizmo PF Member Array Join Thanks 0 Thanked 3 Times in 3 Posts Desktop vs Laptop? Wl like Laptops a lot, Desktops are nice but with the lack of sp Reply With Quote Promote to Article --> 26th January 2011 Jan 2011 Location Republic Of Ireland Posts 190 Reviews Re in 1 Post Without a shadow of a doubt, it's a laptop. Mobile Article --> 30th January 2011 08:58 PM #3 Tony PF Newbie A Reviews Thanks 0 Thanked 0 Times in 0 Posts To each his/h change the parts of my desktop easily. I can do that with a l price for hardware upgrade on laptops is more expensive t traveling then I will use my cellphone so I won't need a lap January 2011 10:19 AM #4 Andyintwitch PF Senior Member Reviews Blog Entries 6 Thanks 0 Thanked 6 Times in 5 Posts knee while watching the tv :-) --> Reply With Quote Promo nerobi10 PF Newbie Array Join Date Jan 2011 Posts 9 Revie
						Mark Forums Read Quick Links Today's Posts View Site Lea Technology Forum Desktops, Laptops, Notebooks &amp; C onclick=prompt('Use the following URL when referencing t blog.','http://www.theproductforum.com/forums/f7/desk About LinkBacks Thread Tools Show Printable Version Ema Search Thread Advanced Search Display Linear Mode Switc January 2011 07:29 PM #1 DarkGizmo PF Member Array Join Thanks 0 Thanked 3 Times in 3 Posts Desktop vs Laptop? Wl

Figure 5.9: Result data set from The Product Forum crawled by Winweb Crawler

5.4. Web Content Extractor

Installation is done by visiting the Newprosoft website under `http://www.newprosoft.com/web-content-extractor.htm`, download the trial version and start the installer.



Related: digital media player tv, digital video player, digital media player 4.3, hd media player, more »

Home > Electronics > MP3 and Digital Media Players > Apple iPod Video 5th Generation White (30 GB) MP3 Player

Apple iPod Video 5th Generation White (30 GB) MP3 Player

159 consumer reviews | Write a Review

Average Rating: Excellent

5 stars	72
4 stars	62
3 stars	17
2 stars	3
1 star	5

Apple iPod 30 GB MP3 Player: A Fantastic Electronic Device

Written: Sep 28 '09

User Rating: Excellent

Sound:
Ease of Use:
Durability:
Battery Life:
Portability:

Pros: awesome sound, easy to use, "on the go" playlists, hard to harm

Cons: battery life is fading after 4 years

The Bottom Line: If you can find a 30GB Video iPod, it's still a solid choice for an MP3 player.

I remember when my husband bought his first iPod. I was completely baffled. Why would you want to have so much music with you at one time; what was the point of that? I could listen to whatever I wanted to at home easily, and I could play multiple CDs in my car. I soon caught on to why he liked his iPod, but didn't feel that I needed one myself. Then for Christmas a few years ago he gave me an **Apple iPod 30 GB MP3 Player**. This was the 5th Generation iPod and the first with video capability. I admit it... despite thinking it was something I would never want or need, I fell in love with it. I'm still using it, almost four later and it is still working well for me.

Selected Data Fields:

Field Name	HTML Path	Type	Max
Positive Feat...	HTML[1].BODY[1].DIV[3].DIV[1].TABLE[1].TBODY[1].TR[1].TD[1].TABLE[1].TBODY[1].TR[1].TD[1].TABLE[1]....	Text	> 25
Negative Fea...	HTML[1].BODY[1].DIV[3].DIV[1].TABLE[1].TBODY[1].TR[1].TD[1].TABLE[1].TBODY[1].TR[1].TD[1].TABLE[1]....	Text	> 25
Main Text	HTML[1].BODY[1].DIV[3].DIV[1].TABLE[1].TBODY[1].TR[1].TD[1].TABLE[1].TBODY[1].TR[1].TD[1].TABLE[1]....	Text	> 25

Field Name: Positive Features
Max Length: > 255
Field Type: Text
Required: No
HTML Path: HTML[1].BODY[1].DIV[3].DIV[1].TABLE[1].TBODY[1].TR[1].TD[1].TABLE[1].TBO
HTML Path 2 (alternative): HTML[1].BODY[1].DIV[3].DIV[1].TABLE[1].TBODY[1].TR[1].TD[1].TABLE[1].TBO
Tag Class:
Tag ID:
Field Text Preview: Pros:awesome sound, easy to use, "on the go" playlists, hard to harm
Cons:battery life is fading after 4 years
The Bottom Line: If you can find a 30GB Video iPod, it's still a solid choice for an M
Text Label: Pros:
Text Processing Script: Sub_String (strText , "Pros:" , vbNewLine , "0")
Script Result: awesome sound, easy to use, "on the go" playlists, hard to harm

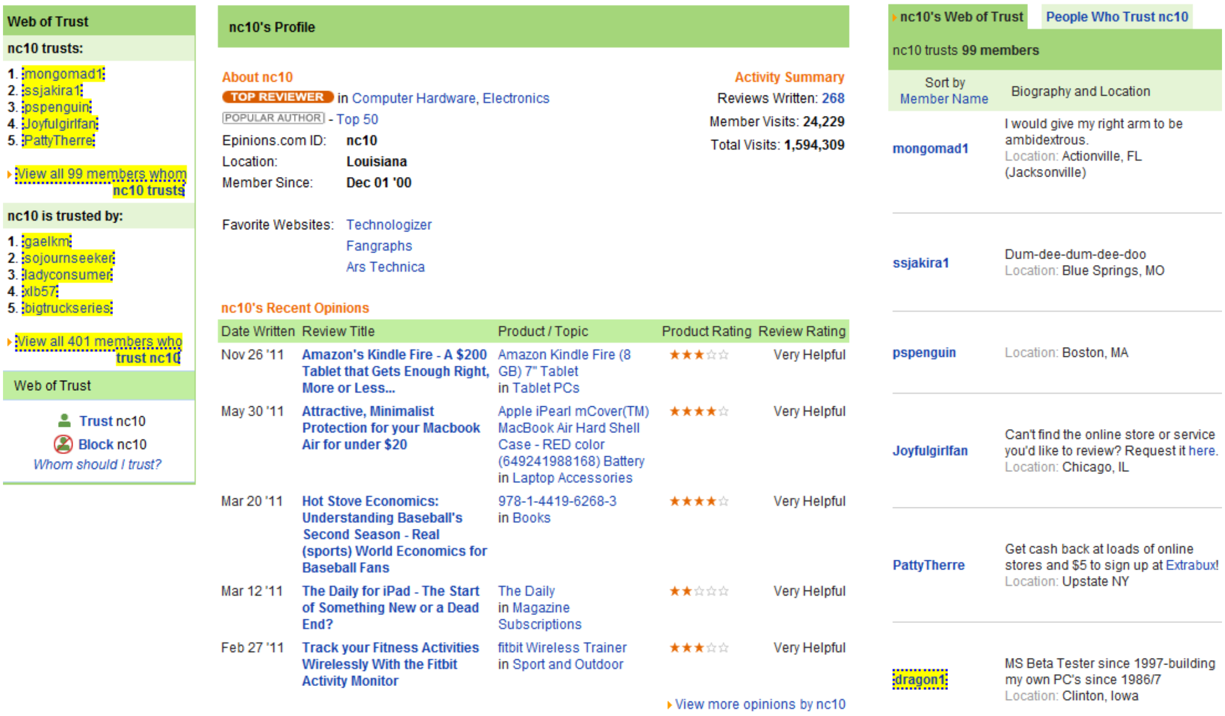
Figure 5.10: Web Content Extractor graphical content selection (left) and target table field specification (right)

Function principle: Web Content Extractor uses an intuitive graphical content and link selection interface where you can specify each link the crawler should follow and the text parts it should extract. You also have to specify the target column field name for each extracted content part, e.g. "body text" for the post message column in a forum. Afterwards the program is able to handle similar sites the same way making it ideal for content extraction on sites that share a similar structure. It is possible to specify the retrieval depth and further to specify differing content extraction patterns for any retrieval level. This enables the user e.g. to specify different extraction patterns for a review overview page and the corresponding detail review sites that open by following the specified links. Besides it is possible to alter the automatic generated extraction script by hand to further refine and debug content selection.

5.4.1. Structure Crawling Epinions

From the review overview starting page the three users of interest were marked manually as links to follow. As second crawl level a reference site from a user with as many incoming and outgoing trust links as possible was selected and the user name patterns were defined on it. Selecting this site was necessary to enable the software to see all possible trust links since the site is created dynamically and shows only so much trust links as the user actually has. For instance, if a user had only 2 trusts-links, only 2 links would show up. Selecting such a site as second level reference and mark the two visible links would result in setting a limit to the crawl process of a maximum of 2 trust links per user. As figure 5.11 shows, even for very active users with a very large number of members trusting him only the first 5 trusted members are shown in the profile. Therefore a third crawl level had to be introduced, that followed the “view all members whom XY trusts” link and crawled all users shown there. Overall two different content extraction patterns had to be defined – one extracting the names on the user profile itself and one extracting the additional names on the “view all members” site – assumed the user has more than five trusting members and the link exist.

It is not possible to merge the results of both sites into one data line because the program handles every page as its own data line with its own column attributes. Hence the program extracted two separate tables.



Web of Trust

nc10 trusts:

1. mongomad1
2. ssjakira1
3. pspenguin
4. Joyfulgirlfan
5. PattyTherre

View all 99 members whom nc10 trusts

nc10 is trusted by:

1. gaelkm
2. sojournseeker
3. ladyconsumer
4. xlb57
5. bigtruckseries

View all 401 members who trust nc10

Web of Trust

Trust nc10

Block nc10

Whom should I trust?

nc10's Profile

About nc10

TOP REVIEWER in Computer Hardware, Electronics

POPULAR AUTHOR - Top 50

Epinions.com ID: **nc10**

Location: **Louisiana**

Member Since: **Dec 01 '00**

Favorite Websites: Technologizer
Fangraphs
Ars Technica

Activity Summary

Reviews Written: 268

Member Visits: **24,229**

Total Visits: **1,594,309**

nc10's Recent Opinions

Date Written	Review Title	Product / Topic	Product Rating	Review Rating
Nov 26 '11	**Amazon's Kindle Fire - A $200 Tablet that Gets Enough Right, More or Less...**	Amazon Kindle Fire (8 GB) 7" Tablet in Tablet PCs	★★★☆☆	Very Helpful
May 30 '11	**Attractive, Minimalist Protection for your Macbook Air for under $20**	Apple iPearl mCover(TM) MacBook Air Hard Shell Case - RED color (649241988168) Battery in Laptop Accessories	★★★★☆	Very Helpful
Mar 20 '11	**Hot Stove Economics: Understanding Baseball's Second Season - Real (sports) World Economics for Baseball Fans**	978-1-4419-6268-3 in Books	★★★★☆	Very Helpful
Mar 12 '11	**The Daily for iPad - The Start of Something New or a Dead End?**	The Daily in Magazine Subscriptions	★★☆☆☆	Very Helpful
Feb 27 '11	**Track your Fitness Activities Wirelessly With the Fitbit Activity Monitor**	fitbit Wireless Trainer in Sport and Outdoor	★★★☆☆	Very Helpful

View more opinions by nc10

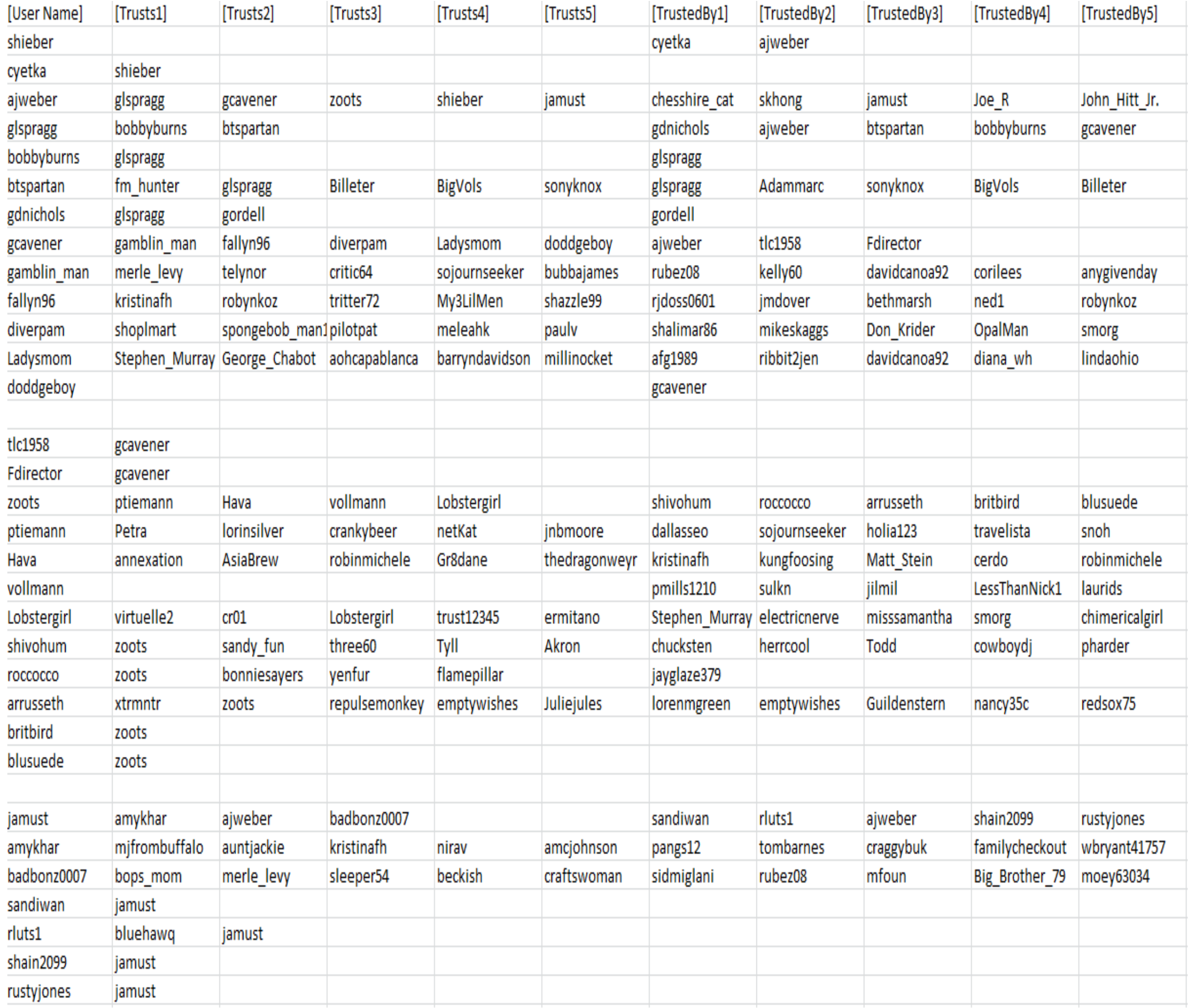
nc10's Web of Trust | People Who Trust nc10

nc10 trusts **99 members**

Sort by Member Name	Biography and Location
mongomad1	I would give my right arm to be ambidextrous. Location: Actionville, FL (Jacksonville)
ssjakira1	Dum-dee-dum-dee-doo Location: Blue Springs, MO
pspenguin	Location: Boston, MA
Joyfulgirlfan	Can't find the online store or service you'd like to review? Request it here. Location: Chicago, IL
PattyTherre	Get cash back at loads of online stores and $5 to sign up at Extrabux! Location: Upstate NY
dragon1	MS Beta Tester since 1997-building my own PC's since 1986/7 Location: Clinton, Iowa

Figure 5.11: User profile (left) and trust detail page (right)

The crawl results were fully satisfactory. Every user and his corresponding trust links could be extracted and stored in tables. Nothing was left out. The only drawback was the fact that Epinions splits greater numbers of trusted members on two pages leading to a result set of two tables. This made manual data merging obligatory afterwards. Since virtually no data mining process works without data preprocessing, this is not a serious flaw.

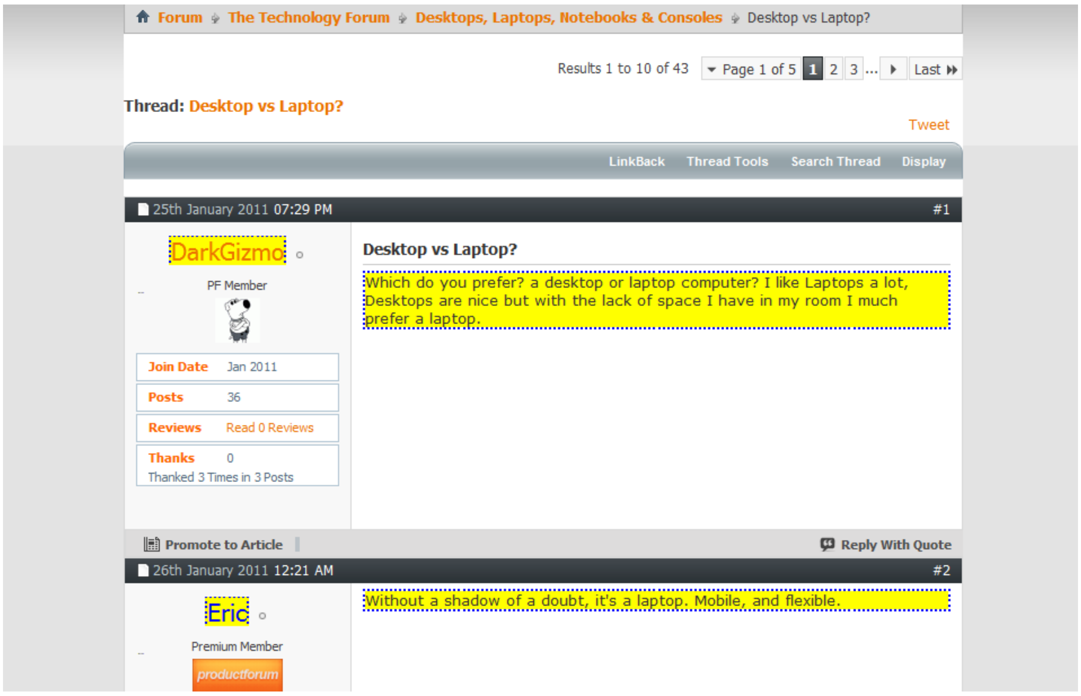


[User Name]	[Trusts1]	[Trusts2]	[Trusts3]	[Trusts4]	[Trusts5]	[TrustedBy1]	[TrustedBy2]	[TrustedBy3]	[TrustedBy4]	[TrustedBy5]
shieber						cyetka	ajweber			
cyetka	shieber									
ajweber	glspragg	gcavener	zoots	shieber	jamust	chesshire_cat	skhong	jamust	Joe_R	John_Hitt_Jr.
glspragg	bobbyburns	btspartan				gdnichols	ajweber	btspartan	bobbyburns	gcavener
bobbyburns	glspragg					glspragg				
btspartan	fm_hunter	glspragg	Billeter	BigVols	sonyknox	glspragg	Adammarc	sonyknox	BigVols	Billeter
gdnichols	glspragg	gordell				gordell				
gcavener	gamblin_man	fallyn96	diverpam	Ladysmom	doddgeboy	ajweber	tlc1958	Fdirector		
gamblin_man	merle_levy	telynor	critic64	sojournseeker	bubbajames	rubez08	kelly60	davidcanoa92	corilees	anygivenday
fallyn96	kristinafh	robynkoz	tritter72	My3LilMen	shazzle99	rjdoss0601	jmdover	bethmarsh	ned1	robynkoz
diverpam	shoplmart	spongebob_man1	pilotpat	meleahk	paulv	shalimar86	mikeskaggs	Don_Krider	OpalMan	smorg
Ladysmom	Stephen_Murray	George_Chabot	aohcapablanca	barryndavidson	millinocket	afg1989	ribbit2jen	davidcanoa92	diana_wh	lindaohio
doddgeboy						gcavener				
tlc1958	gcavener									
Fdirector	gcavener									
zoots	ptiemann	Hava	vollmann	Lobstergirl		shivohum	roccocco	arrusseth	britbird	blusuede
ptiemann	Petra	lorinsilver	crankybeer	netKat	jnbmoore	dallasseo	sojournseeker	holia123	travelista	snoh
Hava	annexation	AsiaBrew	robinmichele	Gr8dane	thedragonweyr	kristinafh	kungfoosing	Matt_Stein	cerdo	robinmichele
vollmann						pmills1210	sulkn	jilmil	LessThanNick1	laurids
Lobstergirl	virtuelle2	cr01	Lobstergirl	trust12345	ermitano	Stephen_Murray	electricnerve	misssamantha	smorg	chimericalgirl
shivohum	zoots	sandy_fun	three60	Tyll	Akron	chucksten	herrcool	Todd	cowboydj	pharder
roccocco	zoots	bonniesayers	yenfur	flamepillar		jayglaze379				
arrusseth	xtrmntr	zoots	repulsemonkey	emptywishes	Juliejules	lorenmgreen	emptywishes	Guildenstern	nancy35c	redsox75
britbird	zoots									
blusuede	zoots									
jamust	amykhar	ajweber	badbonz0007			sandiwan	rluts1	ajweber	shain2099	rustyjones
amykhar	mjfrombuffalo	auntjackie	kristinafh	nirav	amcjohnson	pangs12	tombarnes	craggybuk	familycheckout	wbryant41757
badbonz0007	bops_mom	merle_levy	sleeper54	beckish	craftswoman	sidmiglani	rubez08	mfoun	Big_Brother_79	moey63034
sandiwan	jamust									
rluts1	bluehawq	jamust								
shain2099	jamust									
rustyjones	jamust									

Figure 5.12: Result table 1 gathered from structure crawling on Epinions with Web Content Extractor

5.4.2. Structure Crawling The Product Forum

Starting from the thread overview page any thread link was marked as link the crawler should follow. In the thread the "next" button, showing the next 10 postings, was marked as well on the second crawl level. The extraction pattern defined on the second crawl level consisted of each user name and the content the user posted as figure 5.13 shows.



Thread: Desktop vs Laptop?

25th January 2011 07:29 PM #1

DarkGizmo

PF Member

Join Date Jan 2011
Posts 36
Reviews Read 0 Reviews
Thanks 0
Thanked 3 Times in 3 Posts

Desktop vs Laptop?

Which do you prefer? a desktop or laptop computer? I like Laptops a lot, Desktops are nice but with the lack of space I have in my room I much prefer a laptop.

26th January 2011 12:21 AM #2

Eric

Premium Member

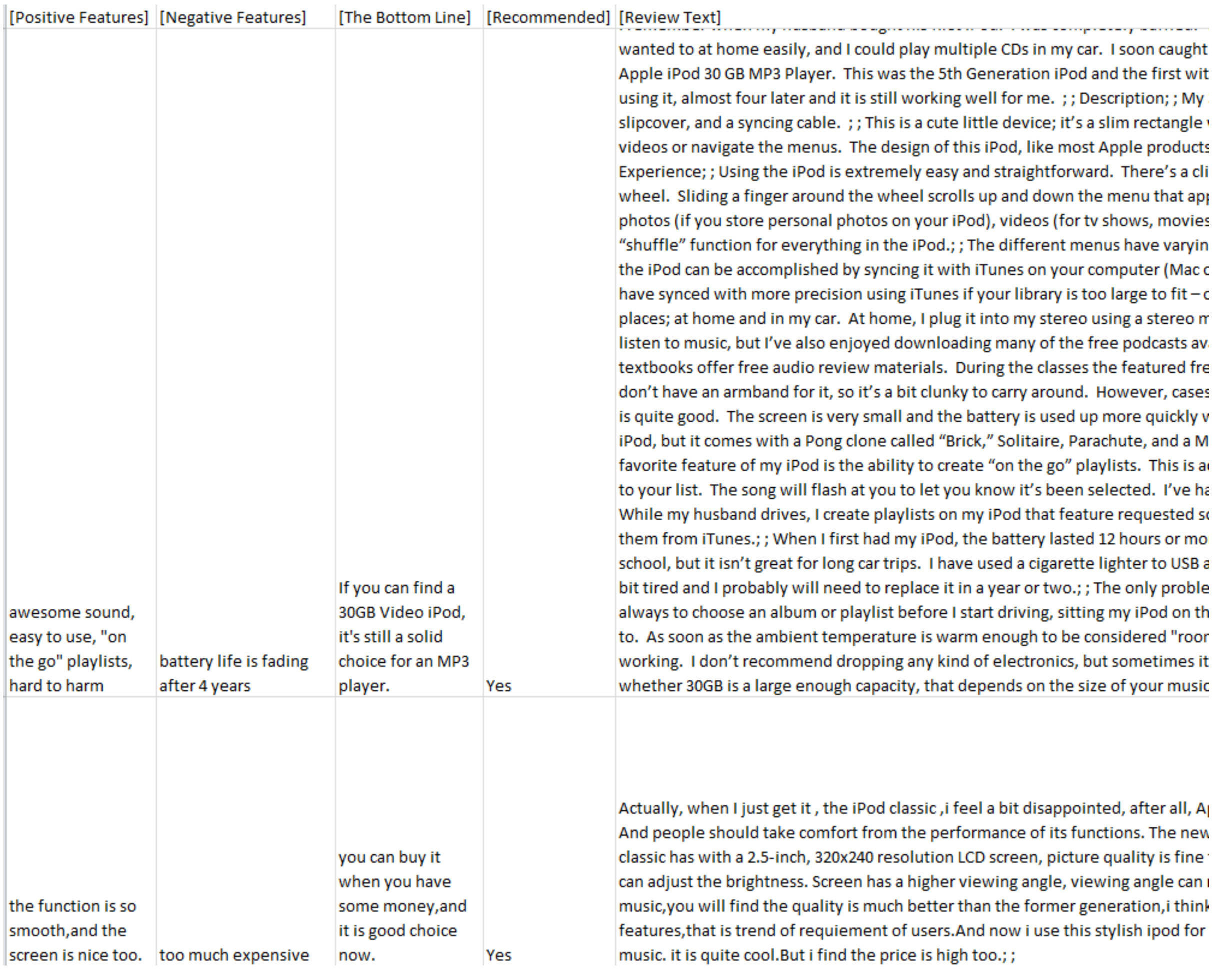
Without a shadow of a doubt, it's a laptop. Mobile, and flexible.

Figure 5.13: Content extraction pattern for structure crawling The Product Forum

The results were satisfactory but had some flaws. First some threads or pages were left out for no obvious reason. Despite manually changing scripts and settings could solve this, a second problem remained in the result table. The table did not contain names of the authors of the second posts in every thread. A third error occurred with the post text. Sometimes the text from author A was shown next to both authors A and B instead of showing the correct text belonging to B.

5.4.3. Content Crawling Epinions

The content extraction was divided in three parts that could be selected in the graphical user interface: "pros", "cons" and "main text". Additionally at the end of each main text field the term "Recommended: Yes (or No)" is placed. Since this is a predetermined feature, it should be separated but the program recognized it as part of the text. Using a self-programmed script everything beginning from the word "Recommended" was excluded from the main text and put in a separate column. This value made later review classification tasks way easier to evaluate since the overall text sentiment is specified with this value.



[Positive Features]	[Negative Features]	[The Bottom Line]	[Recommended]	[Review Text]
awesome sound, easy to use, "on the go" playlists, hard to harm	battery life is fading after 4 years	If you can find a 30GB Video iPod, it's still a solid choice for an MP3 player.	Yes	wanted to at home easily, and I could play multiple CDs in my car. I soon caught Apple iPod 30 GB MP3 Player. This was the 5th Generation iPod and the first wit using it, almost four later and it is still working well for me. ; ; Description; ; My slipcover, and a syncing cable. ; ; This is a cute little device; it's a slim rectangle videos or navigate the menus. The design of this iPod, like most Apple products Experience; ; Using the iPod is extremely easy and straightforward. There's a cli wheel. Sliding a finger around the wheel scrolls up and down the menu that ap photos (if you store personal photos on your iPod), videos (for tv shows, movies "shuffle" function for everything in the iPod.; ; The different menus have varyin the iPod can be accomplished by syncing it with iTunes on your computer (Mac c have synced with more precision using iTunes if your library is too large to fit – c places; at home and in my car. At home, I plug it into my stereo using a stereo m listen to music, but I've also enjoyed downloading many of the free podcasts av textbooks offer free audio review materials. During the classes the featured fre don't have an armband for it, so it's a bit clunky to carry around. However, cases is quite good. The screen is very small and the battery is used up more quickly v iPod, but it comes with a Pong clone called "Brick," Solitaire, Parachute, and a M favorite feature of my iPod is the ability to create "on the go" playlists. This is a to your list. The song will flash at you to let you know it's been selected. I've ha While my husband drives, I create playlists on my iPod that feature requested s them from iTunes.; ; When I first had my iPod, the battery lasted 12 hours or mo school, but it isn't great for long car trips. I have used a cigarette lighter to USB a bit tired and I probably will need to replace it in a year or two.; ; The only proble always to choose an album or playlist before I start driving, sitting my iPod on th to. As soon as the ambient temperature is warm enough to be considered "roor working. I don't recommend dropping any kind of electronics, but sometimes it whether 30GB is a large enough capacity, that depends on the size of your music
the function is so smooth,and the screen is nice too.	too much expensive	you can buy it when you have some money,and it is good choice now.	Yes	Actually, when I just get it , the iPod classic ,i feel a bit disappointed, after all, A And people should take comfort from the performance of its functions. The new classic has with a 2.5-inch, 320x240 resolution LCD screen, picture quality is fine can adjust the brightness. Screen has a higher viewing angle, viewing angle can music,you will find the quality is much better than the former generation,i think features,that is trend of requiement of users.And now i use this stylish ipod for music. it is quite cool.But i find the price is high too.; ;

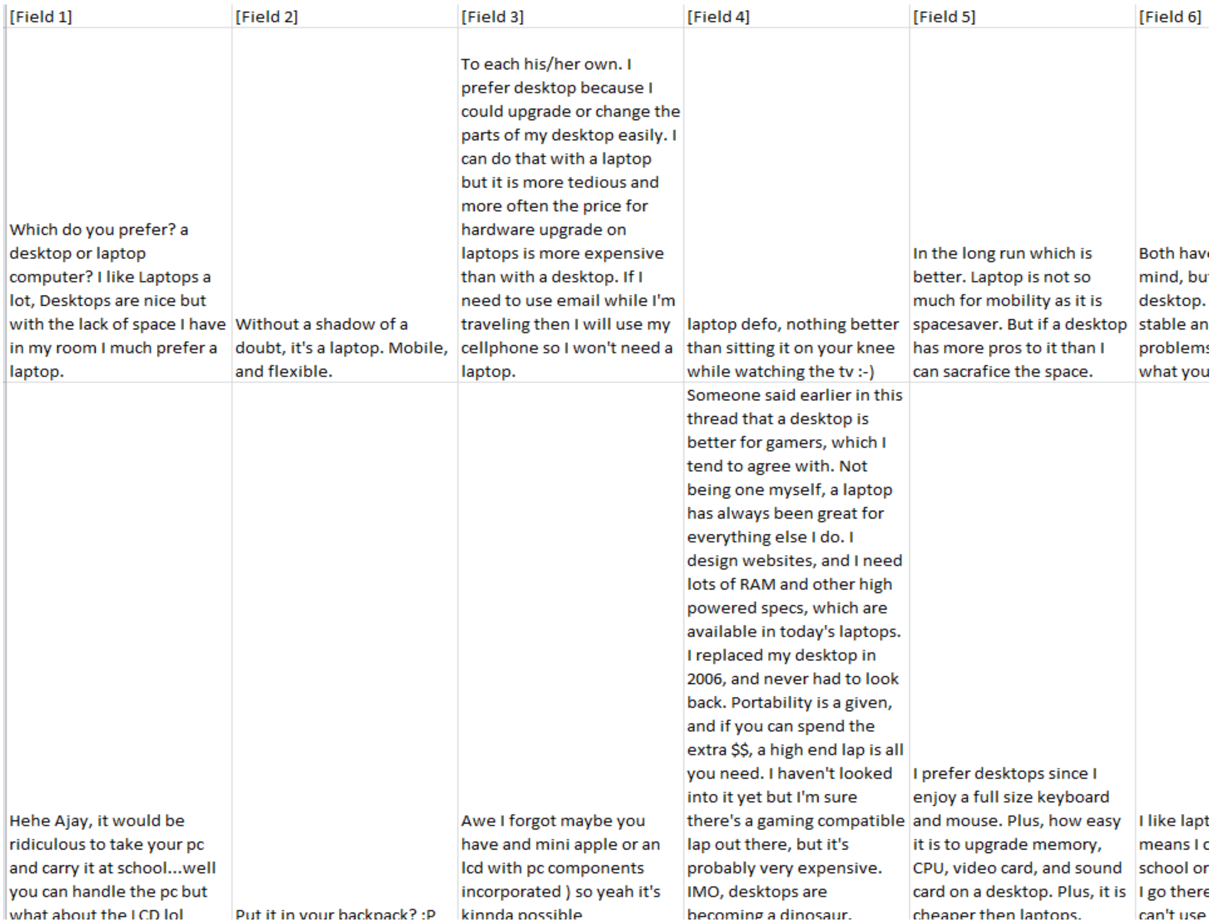
Figure 5.14: Results gathered from content crawling Epinions with Web Content Extractor

Looking at the results nothing can be criticized except few pros and cons were left out when the user made a paragraph in his text. Additionally 9 lines were left out because they exceeded the trial version limit of this software.

5.4.4. Content Crawling The Product Forum

Starting from the "Desktop vs. Laptop" thread site showing off the first 10 posts, the "next" link was specified as follow-up link. Unfortunately the crawler refused to open more than the first three sites and therefore left out the last 12 posts. This could be overcome by creating 5 separate crawl levels – each for every site.

During the content extraction pattern definition process only the post messages were selected since authors have no impact on feature sentiment classification. This resulted in a 10 column wide table where every line represents the 10 postings displayed per page.



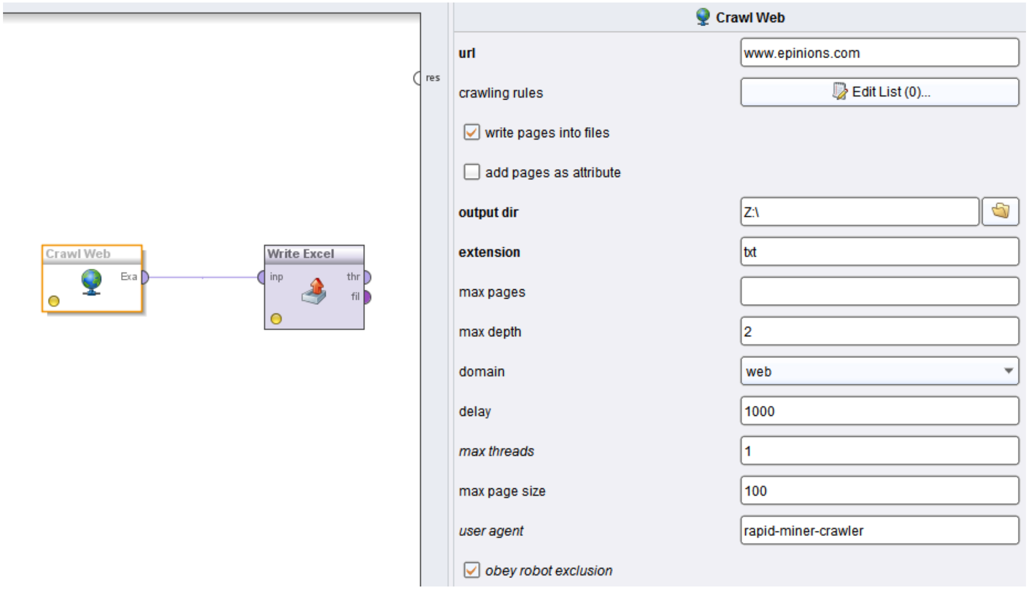
[Field 1]	[Field 2]	[Field 3]	[Field 4]	[Field 5]	[Field 6]
Which do you prefer? a desktop or laptop computer? I like Laptops a lot, Desktops are nice but with the lack of space I have in my room I much prefer a laptop.	Without a shadow of a doubt, it's a laptop. Mobile, and flexible.	To each his/her own. I prefer desktop because I could upgrade or change the parts of my desktop easily. I can do that with a laptop but it is more tedious and more often the price for hardware upgrade on laptops is more expensive than with a desktop. If I need to use email while I'm traveling then I will use my cellphone so I won't need a laptop.	laptop defo, nothing better than sitting it on your knee while watching the tv :-)	In the long run which is better. Laptop is not so much for mobility as it is spacesaver. But if a desktop has more pros to it than I can sacrafice the space.	Both hav mind, bu desktop. stable an problems what you
Hehe Ajay, it would be ridiculous to take your pc and carry it at school...well you can handle the pc but what about the LCD lol	Put it in your backpack? :P	Awe I forgot maybe you have and mini apple or an lcd with pc components incorporated) so yeah it's kinnda possible	Someone said earlier in this thread that a desktop is better for gamers, which I tend to agree with. Not being one myself, a laptop has always been great for everything else I do. I design websites, and I need lots of RAM and other high powered specs, which are available in today's laptops. I replaced my desktop in 2006, and never had to look back. Portability is a given, and if you can spend the extra $$, a high end lap is all you need. I haven't looked into it yet but I'm sure there's a gaming compatible lap out there, but it's probably very expensive. IMO, desktops are becoming a dinosaur.	I prefer desktops since I enjoy a full size keyboard and mouse. Plus, how easy it is to upgrade memory, CPU, video card, and sound card on a desktop. Plus, it is cheaper then laptops.	I like lapt means I c school or I go there can't use

Figure 5.15: Part of the result extracted from The Product Forum with Web Content Extractor

The result is nearly flawless and well usable for the desired mining task. The only two minor flaws were that emoticons were left out and one line was crawled twice. Emoticons could have had a strong sentiment indication value making their absence a loss. Since the duplicate line could easily be recognized and removed manually, this was only a minor flaw.

5.5. Rapid Miner

Installation: The Rapid Miner executable install file can be downloaded from the Rapid-I homepage by following the URL `http://rapid-i.com/content/view/181/190/`. By executing the file the program installs automatically under Windows.



Crawl Web

url	www.epinions.com
crawling rules	Edit List (0)...
☑ write pages into files	
☐ add pages as attribute	
output dir	Z:\
extension	txt
max pages	
max depth	2
domain	web
delay	1000
max threads	1
max page size	100
user agent	rapid-miner-crawler
☑ *obey robot exclusion*	

Figure 5.16: Part of the RapidMiner user interface used for web crawling

Function principle: RapidMiner is a data mining tool that allows graphical workflow modelling using nodes and edges. Each node represents a task while each edge stands for a data flow from one task to another. The simplest way of crawling is to use the “Crawl Web” node and specify it’s parameters in the side-menu like in figure 5.16. Results can be restricted and filtered by using further nodes for (HTML) text filtering. Web crawling is only a small share of functionality that RapidMiner offers. The full functionality is explained in the next chapter where RapidMiner is used as mining tool.

5.5.1. Content Crawling The Product Forum

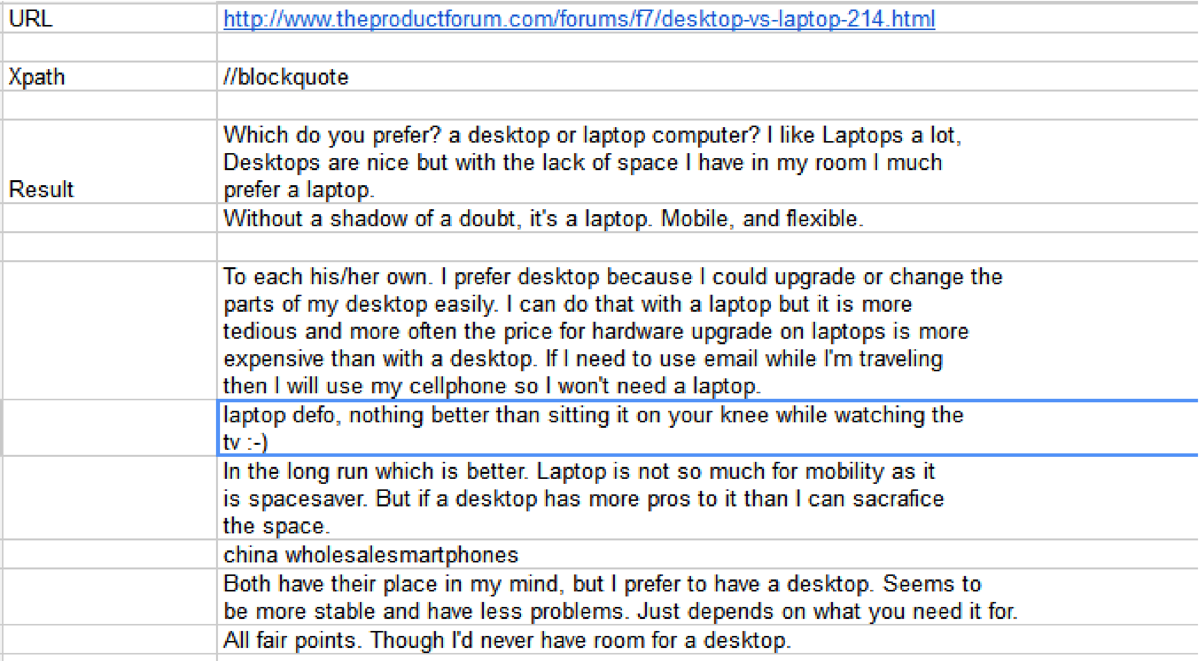
Crawling the whole website extracting source code with the “Crawl Web” node could be done flawlessly. Rapid Miner offers the possibility to extract given URLs and follow-link patterns.

The program refused to follow given links during tests. Program description states that the anchor text should be used as link text to follow but it didn’t work at all. Defining several URLs directly as crawl targets couldn’t get the program to extract more than the first URL either. This made the use of the node “Get Pages” necessary that reads an Excel file containing all URLs. While this method worked fine for this task, the effort for larger projects would be very high.

It is possible to extract information from xml or html files using the “Extract Information” node by defining XPath queries. In theory you should be able to extract very precisely what you want and store it in separate cells where each Xpath query result fills one cell. Unfortunately this did not work in practical use – the process always returned an empty table when typing in the appropriate XPath command. The post messages were stored in the following format:

```
<blockquote class="postcontent restore ">
Which do you prefer? a desktop or laptop computer? I like Laptops a lot,
Desktops are nice but with the lack of space I have in my room I much
prefer a laptop. </blockquote>
```

Commands tested were `//blockquote`, `//div/blockquote`, `//blockquote/text()` and `//blockquote[@class]` and every one of them returned perfect results when crosschecked with google spreadsheet application. This browser based application offers a convenient function to import and process XML documents with XPath commands that can be called with importXML(). The reason why Xpath commands did not work with RapidMiner could not be determined since similar XPath selection commands did not work on other webpages as well.



URL	http://www.theproductforum.com/forums/f7/desktop-vs-laptop-214.html
Xpath	//blockquote
Result	Which do you prefer? a desktop or laptop computer? I like Laptops a lot, Desktops are nice but with the lack of space I have in my room I much prefer a laptop.
	Without a shadow of a doubt, it's a laptop. Mobile, and flexible.
	To each his/her own. I prefer desktop because I could upgrade or change the parts of my desktop easily. I can do that with a laptop but it is more tedious and more often the price for hardware upgrade on laptops is more expensive than with a desktop. If I need to use email while I'm traveling then I will use my cellphone so I won't need a laptop.
	laptop defo, nothing better than sitting it on your knee while watching the tv :-)
	In the long run which is better. Laptop is not so much for mobility as it is spacesaver. But if a desktop has more pros to it than I can sacrafice the space.
	china wholesalesmartphones
	Both have their place in my mind, but I prefer to have a desktop. Seems to be more stable and have less problems. Just depends on what you need it for.
	All fair points. Though I'd never have room for a desktop.

Figure 5.17: Google docs spreadsheet offers decent web content extraction functionality when used on one single page and can be used for XPath test purposes. The figure shows that the XPath command worked correctly.

Another option for extracting information offered by RapidMiner is "String Matching" which allows extracting text between two given strings. Using `blockquote class="postcontent restore ">` as start string and `</blockquote` as end string should have returned the desired text part but it did not return anything. When XPath and String Matching were not used, the crawling result was otherwise flawless as the site source code could be extracted to its full extent.

Since both methods offered for directed content extraction did not work, a third "Extract Content" node had to be used that automatically extracts content from a HTML file. The result was acceptable with storing any post message in a separate cell and adding only moderate noise level as seen in figure 7 below. Still this result needs manual extraction of the desired sentences giving only a small benefit over direct text extraction from the page.

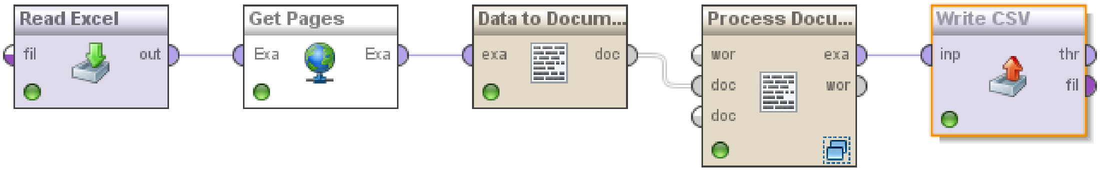
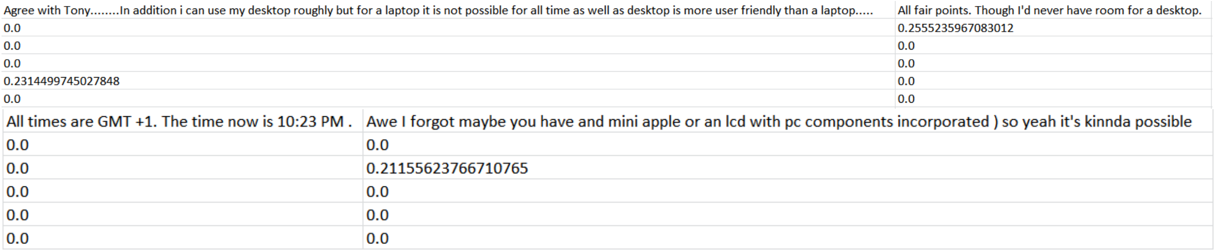


Figure 5.18: Crawling process modelled with RapidMiner. The "Process Documents" node second from the right conatains the "Extract Content" node.



Agree with Tony........In addition i can use my desktop roughly but for a laptop it is not possible for all time as well as desktop is more user friendly than a laptop.....	All fair points. Though I'd never have room for a desktop.
0.0	0.2555235967083012
0.0	0.0
0.0	0.0
0.2314499745027848	0.0
0.0	0.0

All times are GMT +1. The time now is 10:23 PM .	Awe I forgot maybe you have and mini apple or an lcd with pc components incorporated) so yeah it's kinnda possible
0.0	0.0
0.0	0.21155623766710765
0.0	0.0
0.0	0.0
0.0	0.0

Figure 5.19: Crawling result using RapidMiner on The Product Forum. The numbers show TF-IDF for any of the 5 URLs that were crawled.

5.5.2. Structure Crawling The Product Forum

Structure crawling The Product Forum was possible but tedious since, like during content crawling, the crawling link-follow-ups did not work. This time a longer URL list consisting of all 33 pages that can be reached from the thread overview source page at time of testing (October 2012) had to be manually copied from the browser in an Excel file and afterwards they could be processed in the same way as during the content crawling process shown in figure 5.18. The result derived from this process looked comparable to content crawling figure 5.19. The effort necessary for post processing equals the effort for content crawling The Product Forum giving only a small benefit over manual text extraction directly from the website.

5.5.3. Content Crawling Epinions

During content crawling Epinions the same issues showed up as with The Product Forum. The modelled process looked identical to The Product Forum showed in figure 5.18 and neither the link follow-up patterns nor any attempts for XPath data selection or string match filtering seemed to work out.

Another problem was to get sites actually crawled. The only URL on Epinions iPod review sites that actually worked was `http://www.epinions.com/reviews/Apple_iPod_White_30_GB_MA002LL_A_MP3_Player/sec_~opinion_list/pp_~1/pa_~1?sb=1`. All other URLs couldn't be resolved and let the process get stuck at the "Get Pages" node. This is the reason why only the overview page could be crawled that contains only review summaries. The major share of the review text corpora was therefore missing in the final result.

The result is of the same format as The Product Forum showed in figure 5.18.Manual extraction of the desired text part is still necessary.

5.5.4. Structure Crawling Epinions

Structure crawling on Epinions showed the same issues as content crawling. For some reason the crawling node couldn't resolve any site except the reviews overview. Since this fact made crawling the user overview site and its trust-links impossible no results for this task could be obtained at all although the program should theoretically be perfectly suitable to do so.

5.6. Evaluation and Tool Comparison

Commonly used for all crawling tasks is a score from 0 to 5 where 5 is the highest value and 0 means no usable results were achieved at all. The table below as well as later tests show that 5 grades are ideal to put the tools abilities clearly apart. Less would have made smaller differences between tools indistinguishable while more would make each grade too similar to the next one to make a clear distinction. Grading depends on how many of the following three given criteria could be met and to what extent:

- **Automatic link follow-up** means a pattern can be defined that lets the program call the desired pages without further user input.
- A **complete result** means everything shown on the page and relevant for the mining task is there.
- A **clean result** means no noise like HTML tags, duplicate results or irrelevant text from ads, menus or other parts of the site is present in the result table.

☹	No (usable) results could be achieved
☺	None of the three criteria could be met to their full extent. Some result could be achieved but it misses crucial parts and is noisy.
☺☺	Only one criterion could be met. Mediocre result quality either missing important parts or containing much noise.
☺☺☺	Two criteria were met almost to their full extent. The tool provides either good result quality with failed link follow-up or mediocre result quality with working link follow-up.
☺☺☺☺	All three criteria could be met to a large extent. Very good result with only few errors and little post processing needed. Link follow-ups work.
☺☺☺☺☺	Flawless result with no content and only minor formal errors. Link follow-ups work and only minimal post processing is needed.

Table 5.3: Scale used to grade tools.

Tool	**Newprosoft Web Content Extractor**	**winweb Crawler v 2.0**	**RapidMiner**
Homepage	`http://www.newprosoft.com/web-content-extractor.htm`	`http://www.winwebcrawler.com/`	`http://rapid-i.com/content/view/181/190/`
Structure Crawl Epinions	☺☺☺☺☺	☹	☹
Automatic link follow-up worked	Yes	No	No
Result complete	Yes, due to site structure issues more than 5 trusted users had to be extracted in a separate table.	No user profile site could be read and no results could be obtained.	No user profile site could be read and no results could be obtained.
Result is "clean"	Yes	No	No

Table 5.4: Tool crawling functionality compared in practical use part 1

Tool	Newprosoft Web Content Extractor	winweb Crawler v 2.0	RapidMiner
Content Crawl Epinions	☺☺☺☺☺	☺☺	☺
Automatic link follow-up worked	Yes	No, every detail review page link had to be specified manually.	No, neither link follow-ups nor a specified link list did work.
Result complete	Yes	Yes	No, only the overview site could be crawled.
Result is "clean"	Yes	No	No
Structure Crawl The Product Forum	☺☺☺	☺☺	☺☺
Automatic link follow-up worked	Yes	Mostly yes but some sites were left out.	No, a manually created link list had to be used
Result complete	No, some threads or pages were left out for no obvious reason. Some member names could not be resolved	No, threads with more than 10 posts were not extracted completely.	Yes
Result is "clean"	Mostly yes but sometimes one thread was crawled two times. Otherwise clean result.	No, data is very noisy including HTML tags and text surrounding the post messages.	No
Content Crawl The Product Forum	☺☺☺☺	☺☺☺	☺☺
Automatic link follow-up worked	Yes	Yes	No, a manually created link list had to be used.
Result complete	Yes, but emoticons were left out	Yes	Yes
Result is "clean"	Mostly yes but partially duplicate crawls	No, very noisy data and duplicate data sets.	No, it includes noise.

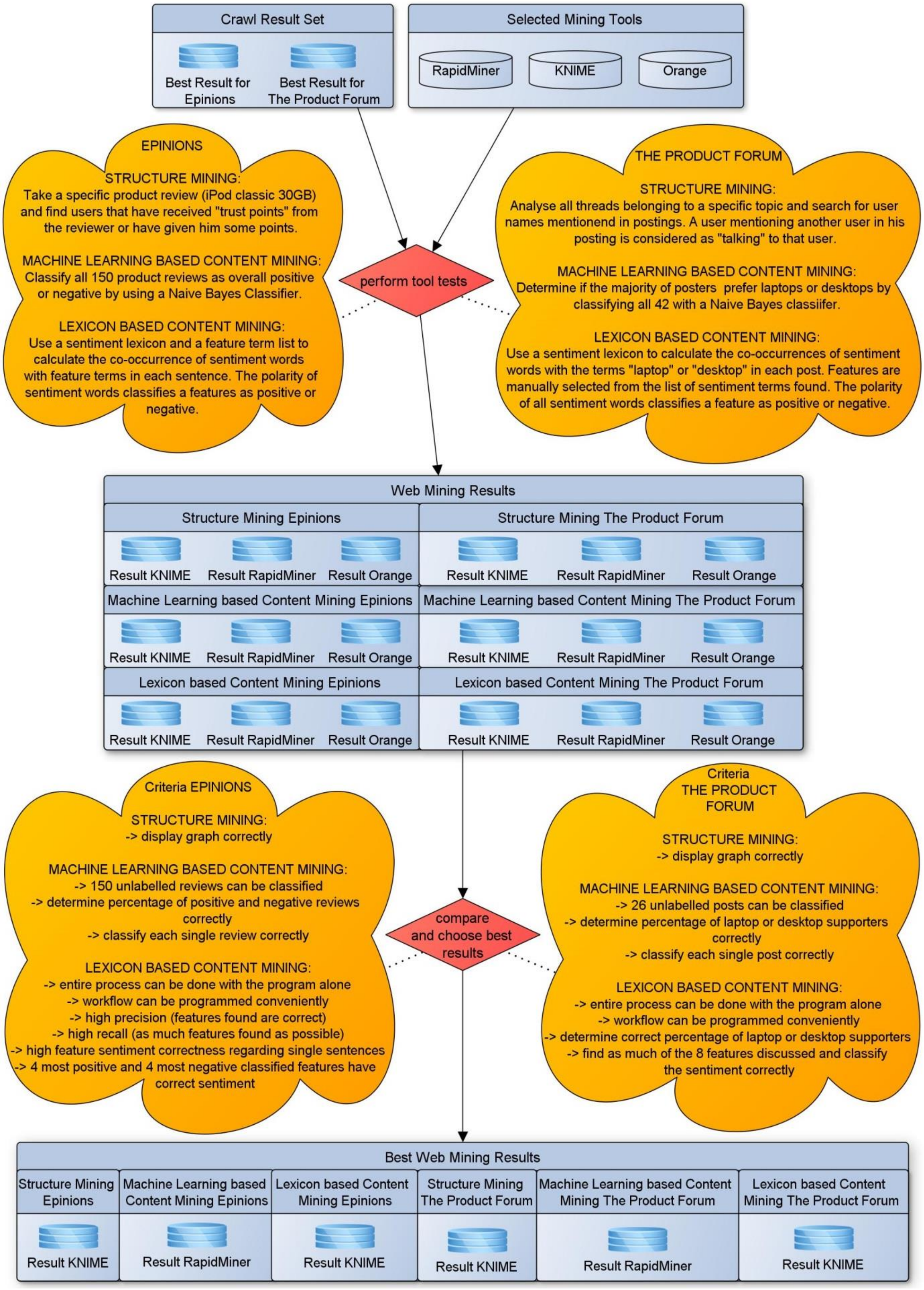
Table 5.5: Tool crawling functionality compared in practical use part 2

5.6.1. Choosing the best Results as Database for further Data Mining

The tool comparison table clearly shows which tool delivered the best results. Newprosofts Web Content Extractor was the only tool in the field that could deliver satisfactory results for all four crawling disciplines. For structure crawling Epinions it was the only tool to gather any of the desired data at all. Since the tool beats the other tools in every single crawl discipline tested, it is clearly the best suited tool for web crawling in both online communities tested.

Consequently, the following mining tasks use those result tables that were created during all four crawling processes using Web Content Extractor. All results gathered with other tools are not further considered.

6. Mining Tool Tests


Crawl Result Set
Best Result for Epinions
Best Result for The Product Forum
Selected Mining Tools
RapidMiner
KNIME
Orange
perform tool tests
Web Mining Results
Structure Mining Epinions
Result KNIME
Result RapidMiner
Result Orange
Structure Mining The Product Forum
Result KNIME
Result RapidMiner
Result Orange
Machine Learning based Content Mining Epinions
Result KNIME
Result RapidMiner
Result Orange
Machine Learning based Content Mining The Product Forum
Result KNIME
Result RapidMiner
Result Orange
Lexicon based Content Mining Epinions
Result KNIME
Result RapidMiner
Result Orange
Lexicon based Content Mining The Product Forum
Result KNIME
Result RapidMiner
Result Orange
compare and choose best results
Best Web Mining Results
Structure Mining Epinions
Result KNIME
Machine Learning based Content Mining Epinions
Result RapidMiner
Lexicon based Content Mining Epinions
Result KNIME
Structure Mining The Product Forum
Result KNIME
Machine Learning based Content Mining The Product Forum
Result RapidMiner
Lexicon based Content Mining The Product Forum
Result KNIME

EPINIONS

STRUCTURE MINING:
Take a specific product review (iPod classic 30GB) and find users that have received "trust points" from the reviewer or have given him some points.

MACHINE LEARNING BASED CONTENT MINING:
Classify all 150 product reviews as overall positive or negative by using a Naive Bayes Classifier.

LEXICON BASED CONTENT MINING:
Use a sentiment lexicon and a feature term list to calculate the co-occurrence of sentiment words with feature terms in each sentence. The polarity of sentiment words classifies a features as positive or negative.

THE PRODUCT FORUM

STRUCTURE MINING:
Analyse all threads belonging to a specific topic and search for user names mentionend in postings. A user mentioning another user in his posting is considered as "talking" to that user.

MACHINE LEARNING BASED CONTENT MINING:
Determine if the majority of posters prefer laptops or desktops by classifying all 42 with a Naive Bayes classiifer.

LEXICON BASED CONTENT MINING:
Use a sentiment lexicon to calculate the co-occurrences of sentiment words with the terms "laptop" or "desktop" in each post. Features are manually selected from the list of sentiment terms found. The polarity of all sentiment words classifies a feature as positive or negative.

Criteria EPINIONS

STRUCTURE MINING:
-> display graph correctly

MACHINE LEARNING BASED CONTENT MINING:
-> 150 unlabelled reviews can be classified
-> determine percentage of positive and negative reviews correctly
-> classify each single review correctly

LEXICON BASED CONTENT MINING:
-> entire process can be done with the program alone
-> workflow can be programmed conveniently
-> high precision (features found are correct)
-> high recall (as much features found as possible)
-> high feature sentiment correctness regarding single sentences
-> 4 most positive and 4 most negative classified features have correct sentiment

Criteria THE PRODUCT FORUM

STRUCTURE MINING:
-> display graph correctly

MACHINE LEARNING BASED CONTENT MINING:
-> 26 unlabelled posts can be classified
-> determine percentage of laptop or desktop supporters correctly
-> classify each single post correctly

LEXICON BASED CONTENT MINING:
-> entire process can be done with the program alone
-> workflow can be programmed conveniently
-> determine correct percentage of laptop or desktop supporters
-> find as much of the 8 features discussed and classify the sentiment correctly

Figure 6.1: Overview of the web mining process performed in this chapter. Each number in a white rectangle describes the chapters where the corresponding task or result is described in further detail.

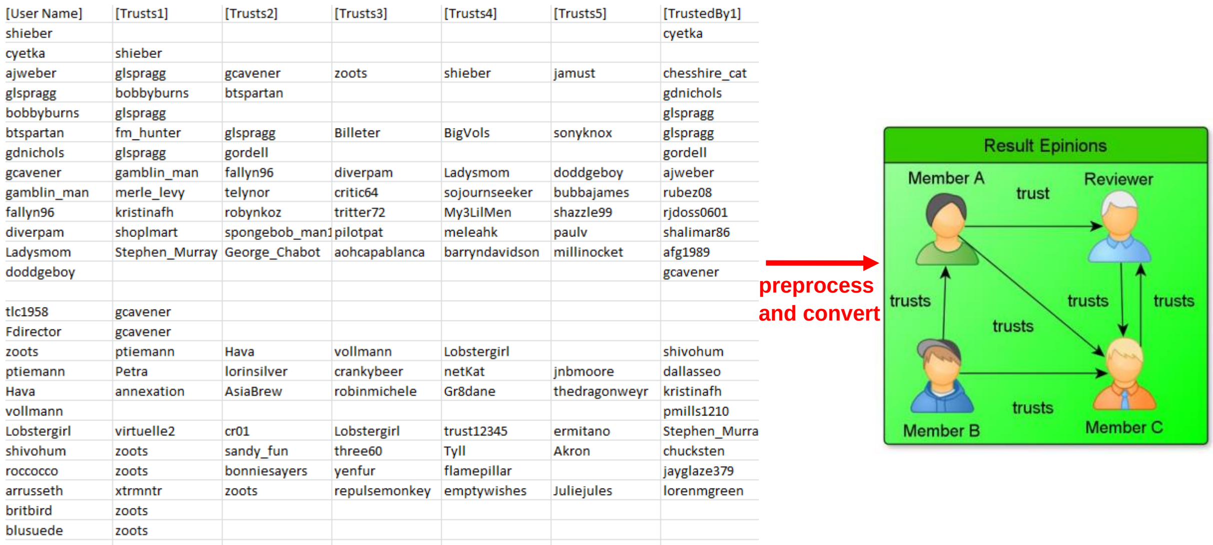
6.1. General Mining Process and desired Results

Since the best crawling results could be obtained by using Web Content Extractor, those results were used for the following data mining process. As with the crawling result presentation in chapter 5 this chapter starts with a general introduction in the mining process as it was done with every tool in the same basic way just with differing methods and result variations.

As with the crawling process the following two subchapters give a general overview of the ideal crawling process results and a score of 0 to 5 will be given for any of the tools under review and displayed in the comparison table in chapter 6.5.

6.1.1. Structure Mining Epinions

Structure mining on Epinions started with the raw data crawled with the Web Content Extractor that looked like shown in chapter 5.4.1., figure 5.10. As this figure shows each data line starts with the user followed by users this user trusts and users this user is trusted by. The data mining process has to convert the table in a format processable by mining software and display users this user trusts as outgoing links and users this user is trusted by as incoming links. The task is therefore to convert the output table from Web Content Extractor in a graph as shown in figure 6.2. Since the vast number of nodes in the structure mining tasks makes the graph large and confusing, different views and detail extraction is important to make desired results visible to a human spectator as shown in figures 6.3 and 6.4.



[User Name]	[Trusts1]	[Trusts2]	[Trusts3]	[Trusts4]	[Trusts5]	[TrustedBy1]
shieber						cyetka
cyetka	shieber					
ajweber	glspragg	gcavener	zoots	shieber	jamust	chesshire_cat
glspragg	bobbyburns	btspartan				gdnichols
bobbyburns	glspragg					glspragg
btspartan	fm_hunter	glspragg	Billeter	BigVols	sonyknox	glspragg
gdnichols	glspragg	gordell				gordell
gcavener	gamblin_man	fallyn96	diverpam	Ladysmom	doddgeboy	ajweber
gamblin_man	merle_levy	telynor	critic64	sojournseeker	bubbajames	rubez08
fallyn96	kristinafh	robynkoz	tritter72	My3LilMen	shazzle99	rjdoss0601
diverpam	shoplmart	spongebob_man1	pilotpat	meleahk	paulv	shalimar86
Ladysmom	Stephen_Murray	George_Chabot	aohcapablanca	barryndavidson	millinocket	afg1989
doddgeboy						gcavener
tlc1958	gcavener					
Fdirector	gcavener					
zoots	ptiemann	Hava	vollmann	Lobstergirl		shivohum
ptiemann	Petra	lorinsilver	crankybeer	netKat	jnbmoore	dallasseo
Hava	annexation	AsiaBrew	robinmichele	Gr8dane	thedragonweyr	kristinafh
vollmann						pmills1210
Lobstergirl	virtuelle2	cr01	Lobstergirl	trust12345	ermitano	Stephen_Murra
shivohum	zoots	sandy_fun	three60	Tyll	Akron	chucksten
roccocco	zoots	bonniesayers	yenfur	flamepillar		jayglaze379
arrusseth	xtrmntr	zoots	repulsemonkey	emptywishes	Juliejules	lorenmgreen
britbird	zoots					
blusuede	zoots					

Figure 6.2: Transformation from crawling output table to the desired user trust relationship graph

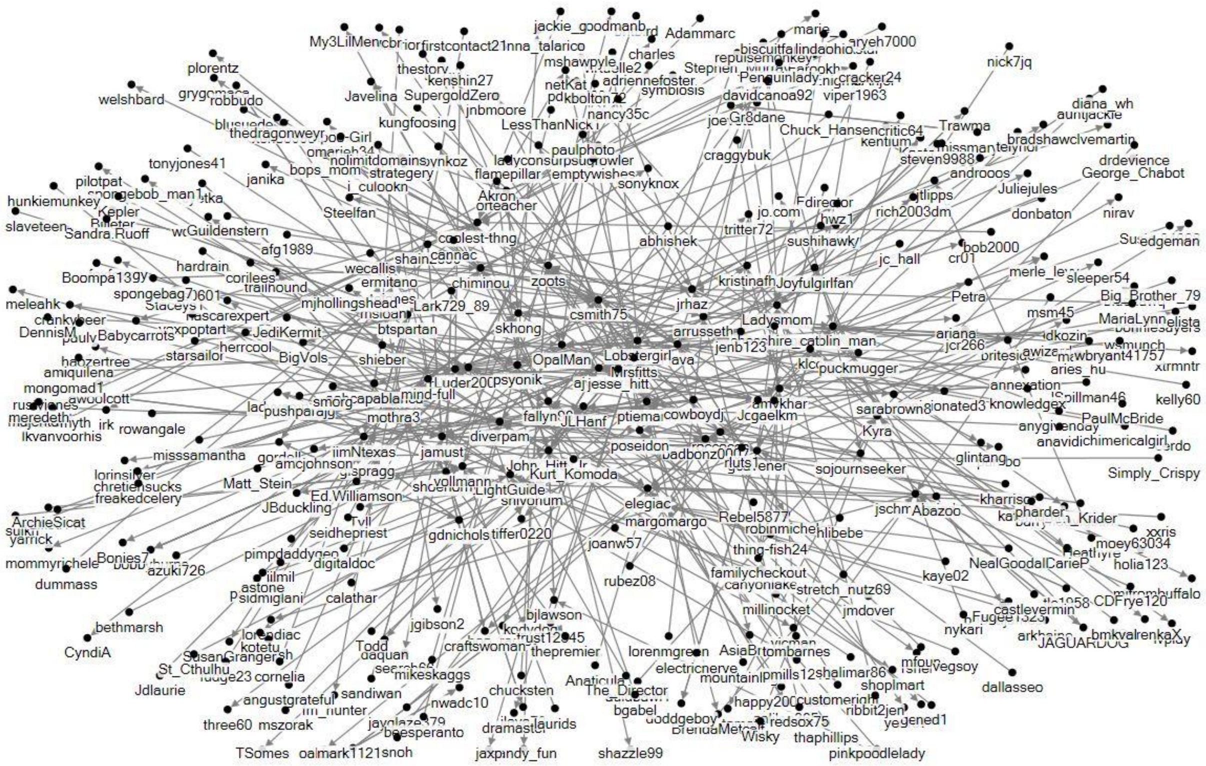


Figure 6.3: The unsorted Epinions user trust result graph created with NodeXL is a mess. It is nearly impossilbe to see any information by looking on it.

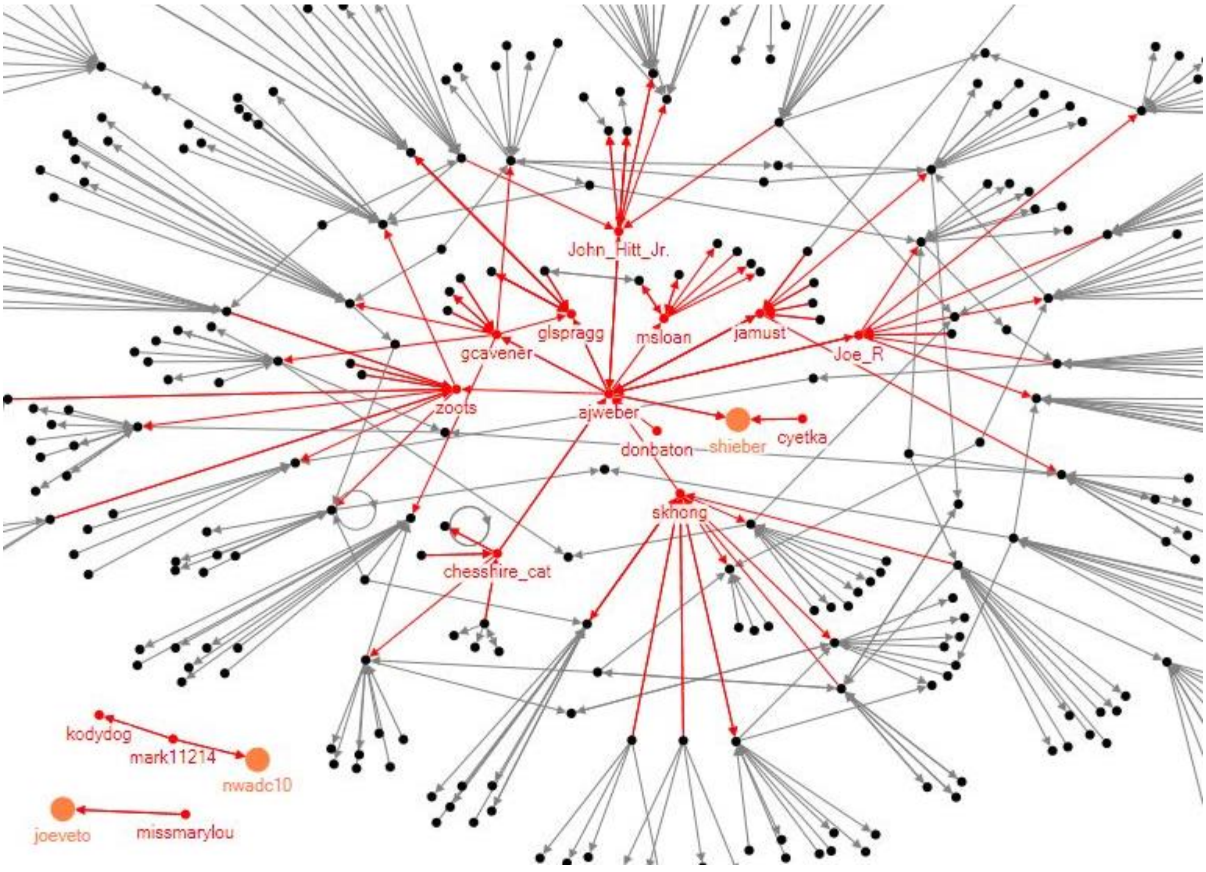


Figure 6.4: Reference result graph: While still not perfect this graph using the Harel-Koren Fast Multiscale Edge aligning algorithm and manual alignment correction as well as colorizing users under review as well as their neighbors shows well which connections exist between the users of interest.

The unclear edge alignment in figure 6.3 makes the importance of highlighting or extracting nodes of special interest obvious. Figure 6.4 shows the same graph with the edges aligned by the so named Harel-Koren Fast Multiscale Edge aligning algorithm[31]. Besides only those user names are displayed and their nodes and corresponding edges are hilighted red, that are two or less nodes away from the three main users shieber, nwadc10 and joeveto. This means any highlighted user will either trust or is trusted by one of these three users directly or he trusts or is trusted by a user that trusts one of these three users.

According to the graph in figure 6.4, mwadc10 and joeveto are just very loosely tied to the community meaning they got trust only from relatively inactive users. Shieber got trust from ajweber which clearly is the most active user under review and most likely a more trustworthy user. No suspicious links can be detected between the users. Overall it can be said that negative 1-star are mainly given by inexperienced users.

This task is one of the easier compared to other mining tasks that require content mining and does not necessarily rely on data mining software but can be solved, for instance, with the spreadsheet program Excel combined with the extension NodeXL [49] as well. Anyways it is interesting to know if the different data mining programs are capable of solving this kind of table transformation and graph visualization task as well.

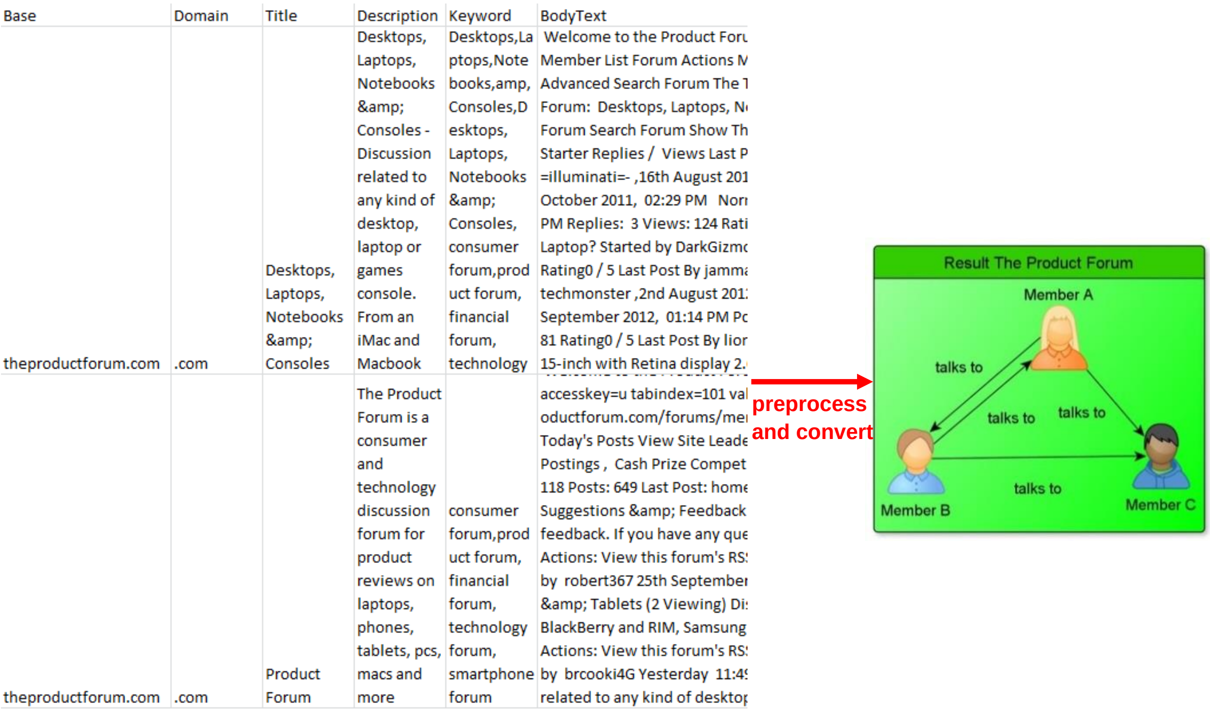
Evaluation of the result graph was done by measuring the result error rate. This rate is a numeric value counting any deviance in the result graph compared to the reference graph in figure 6.4. Deviances can be a missing or additional node or edge. In addition any node that cannot be seen because it is more than half masked by other objects as well as any hidden edge will be counted as 1 error point. Any possibilities for node arrangement optimizations are utilized to its full extent. The overall clearness of graphical representation will not be counted in this evaluation value as it is mainly subjective. Instead it will be subjectively evaluated and considered in the total evaluation score.

6.1.2. Structure Mining The Product Forum

Structure mining The Product Forum works very similar to Epinions. The major difference lies in the node source format. On Epinions the edges alias the user names (the nodes) could be crawled directly from the site link text. On The Product Forum only the authors that form the source edges are in its own data cells while the target edges are hidden in plain text from the forum post messages. The source table format as well as the resulting graph type are shown in figure 6.5.

This requires the additional step of finding and extracting user names from those messages. Determining which terms are user names was done by using the post author names as references. While manual review has shown that sometimes users were mentioned that did not take part in the discussion in this thread, extracting those names would require further crawling on the website and therefore was excluded from this analysis. For test purposes it is fully sufficient to limit the result graph on those users that actively took part on the discussion.

[31] An algorithm that hierarchically orders nodes in tree-form. For further information see [48]



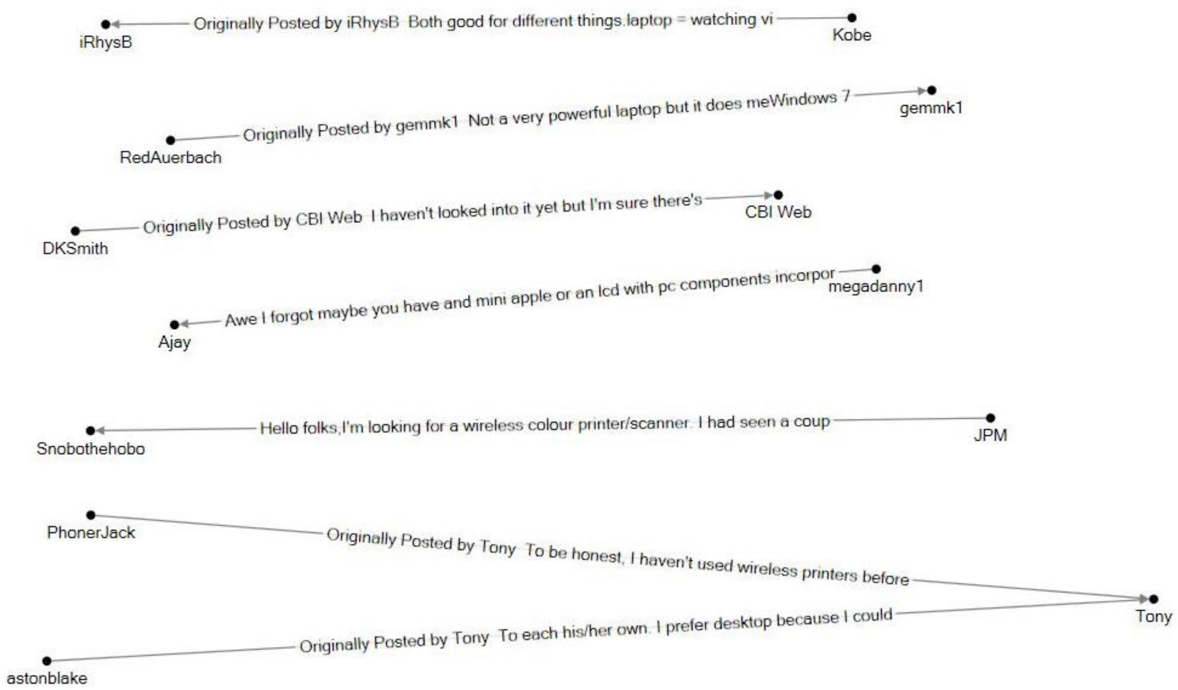
Base	Domain	Title	Description	Keyword	BodyText
theproductforum.com	.com	Desktops, Laptops, Notebooks &amp; Consoles	Desktops, Laptops, Notebooks &amp; Consoles - Discussion related to any kind of desktop, laptop or games console. From an iMac and Macbook	Desktops,Laptops,Notebooks,amp,Consoles,Desktops, Laptops, Notebooks &amp; Consoles, consumer forum,product forum, financial forum, technology	Welcome to the Product Foru Member List Forum Actions M Advanced Search Forum The T Forum: Desktops, Laptops, N Forum Search Forum Show Th Starter Replies / Views Last P =illuminati=- ,16th August 201 October 2011, 02:29 PM Nor PM Replies: 3 Views: 124 Rati Laptop? Started by DarkGizm Rating0 / 5 Last Post By jamm techmonster ,2nd August 201 September 2012, 01:14 PM P 81 Rating0 / 5 Last Post By lio 15-inch with Retina display 2.
theproductforum.com	.com	Product Forum	The Product Forum is a consumer and technology discussion forum for product reviews on laptops, phones, tablets, pcs, macs and more	consumer forum,product forum, financial forum, technology forum, smartphone forum	accesskey=u tabindex=101 val oductforum.com/forums/me Today's Posts View Site Leade Postings , Cash Prize Compet 118 Posts: 649 Last Post: hom Suggestions &amp; Feedback feedback. If you have any que Actions: View this forum's RS by robert367 25th Septembe &amp; Tablets (2 Viewing) Di BlackBerry and RIM, Samsung Actions: View this forum's RS by brcooki4G Yesterday 11:4 related to any kind of deskto

Figure 6.5: Transformation from crawling output table to the member contact graph

The result is of the same type as structure mining on Epinions. Since the number of result edges is significantly lower, highlighting edges and aligning them in a more clearly arranged way is not necessary. Additionally the ability to display post messages as edge labels is tested during this structure mining run.

The manually created reference graph in figure 6.6 shows that forum members are rather loosely connected to each other. Manual review of postings confirmed that. The Product Forum postings are relatively short in general and only in few cases users respond to a specific post mentioning the name or quoting the original message. Edge labels show to be insufficient to display the whole post message but at least they give a rough idea of the post content. Any message beginning with “Originally Posted by iRhysB” is a quoted message where the post author directly responds to a message written by the target user.

Evaluation was done in the same way as for structure mining Epinions. The more correct nodes and edges the better the result. In addition the ability to display edge labels slightly affects the result.



iRhysB
Originally Posted by iRhysB Both good for different things.laptop = watching vi
Kobe
Originally Posted by gemmk1 Not a very powerful laptop but it does meWindows 7
gemmk1
RedAuerbach
Originally Posted by CBI Web I haven't looked into it yet but I'm sure there's
CBI Web
DKSmith
Awe I forgot maybe you have and mini apple or an lcd with pc components incorpor
megadanny1
Ajay
Snobothehobo
Hello folks,I'm looking for a wireless colour printer/scanner. I had seen a coup
JPM
PhonerJack
Originally Posted by Tony To be honest, I haven't used wireless printers before
Tony
Originally Posted by Tony To each his/her own. I prefer desktop because I could
astonblake

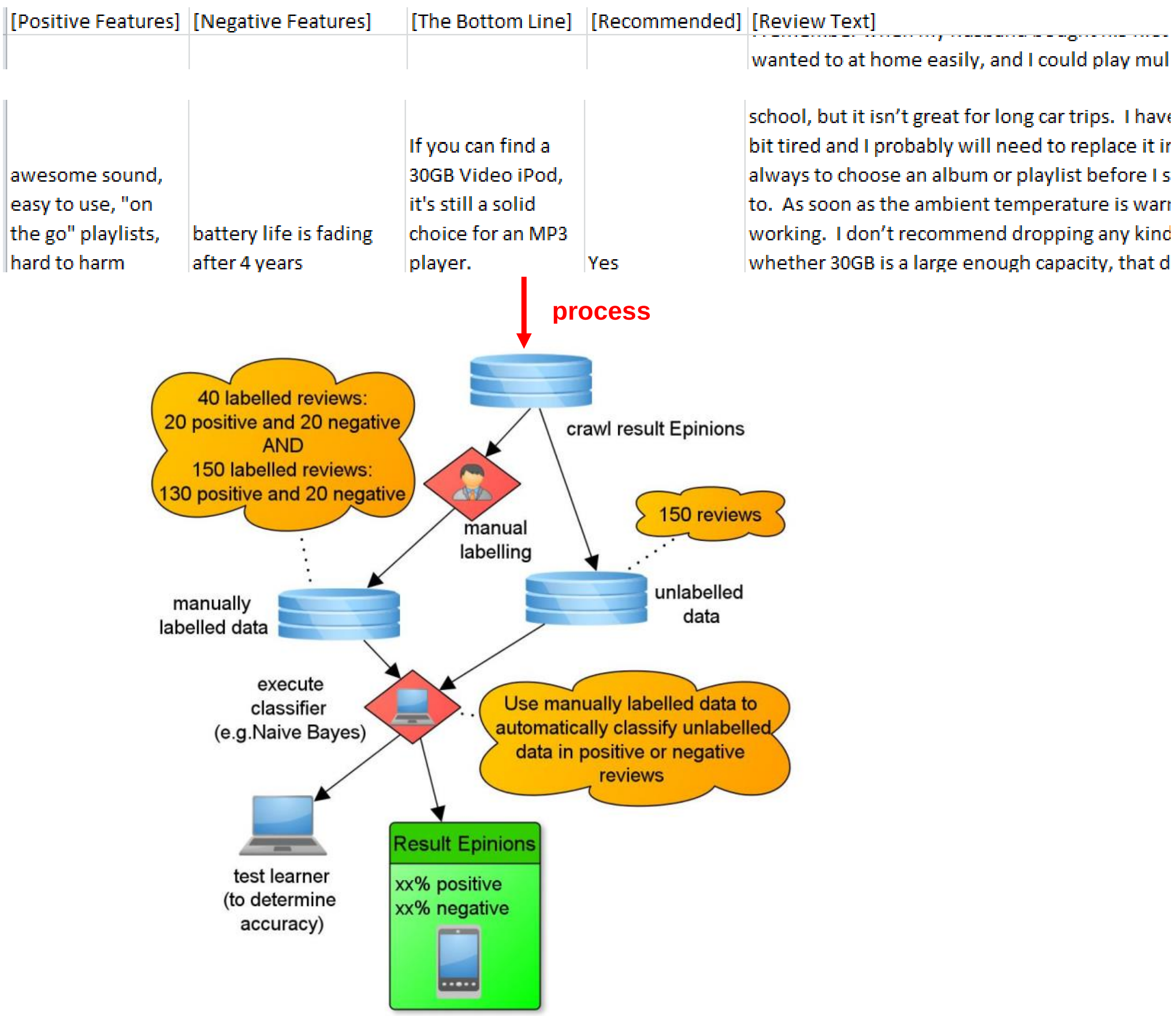
Figure 6.6: Reference result graph: The Product Forum Structure graph created with NodeXL shows there are only few messages directly addressing another user in this forum.

6.1.3. Machine Learning based Content Mining Epinions

As mentioned in chapter 4.3, content mining can be done either with a lexicon based or a text classification model learning approach. Since mining tools can be able to do one or both of those methods the test procedure will go through one or two content mining passes.

Using the text classification approach allows checking if a whole review is overall positive or negative without giving details about the reason why it is so or the feature that is regarded positive or negative. Since Epinions already offers a “recommended” field where the reviewer can choose from two options “recommended: yes” or “no” this mining task does not reveal any additional knowledge. For test purposes as this thesis requires it is perfectly suitable since you instantly know if the mining tool classified the review text correctly or wrong by comparing the tool classification result with the user given recommendation value saving effort for manual text reviews.

Using the text corpora from the crawling result table, a varying number of review text cells is manually labelled and builds the training example for automatically classifying the rest of the remaining reviews. Tests are performed with 40 labelled examples from a total number of 150 data sets whereas one half of the examples are positive and the other half negative. Classification is done once on the 40 examples and once on the whole 150 reviews. In addition classification of all reviews using all 150 labelled reviews is tested too. As classifier Naive Bayes or Support Vector Machines are used. The process is graphically explained in figure 6.7.



[Positive Features]	[Negative Features]	[The Bottom Line]	[Recommended]	[Review Text]
				wanted to at home easily, and I could play mul
awesome sound, easy to use, "on the go" playlists, hard to harm	battery life is fading after 4 years	If you can find a 30GB Video iPod, it's still a solid choice for an MP3 player.	Yes	school, but it isn't great for long car trips. I have bit tired and I probably will need to replace it ir always to choose an album or playlist before I s to. As soon as the ambient temperature is warı working. I don't recommend dropping any kinc whether 30GB is a large enough capacity, that d

process
40 labelled reviews: 20 positive and 20 negative AND 150 labelled reviews: 130 positive and 20 negative
crawl result Epinions
manual labelling
150 reviews
manually labelled data
unlabelled data
execute classifier (e.g.Naive Bayes)
Use manually labelled data to automatically classify unlabelled data in positive or negative reviews
test learner (to determine accuracy)
Result Epinions
xx% positive
xx% negative

Figure 6.7: Transformation from the crawling output table to the desired positive and negative review distribution using supervised machine learning techniques.

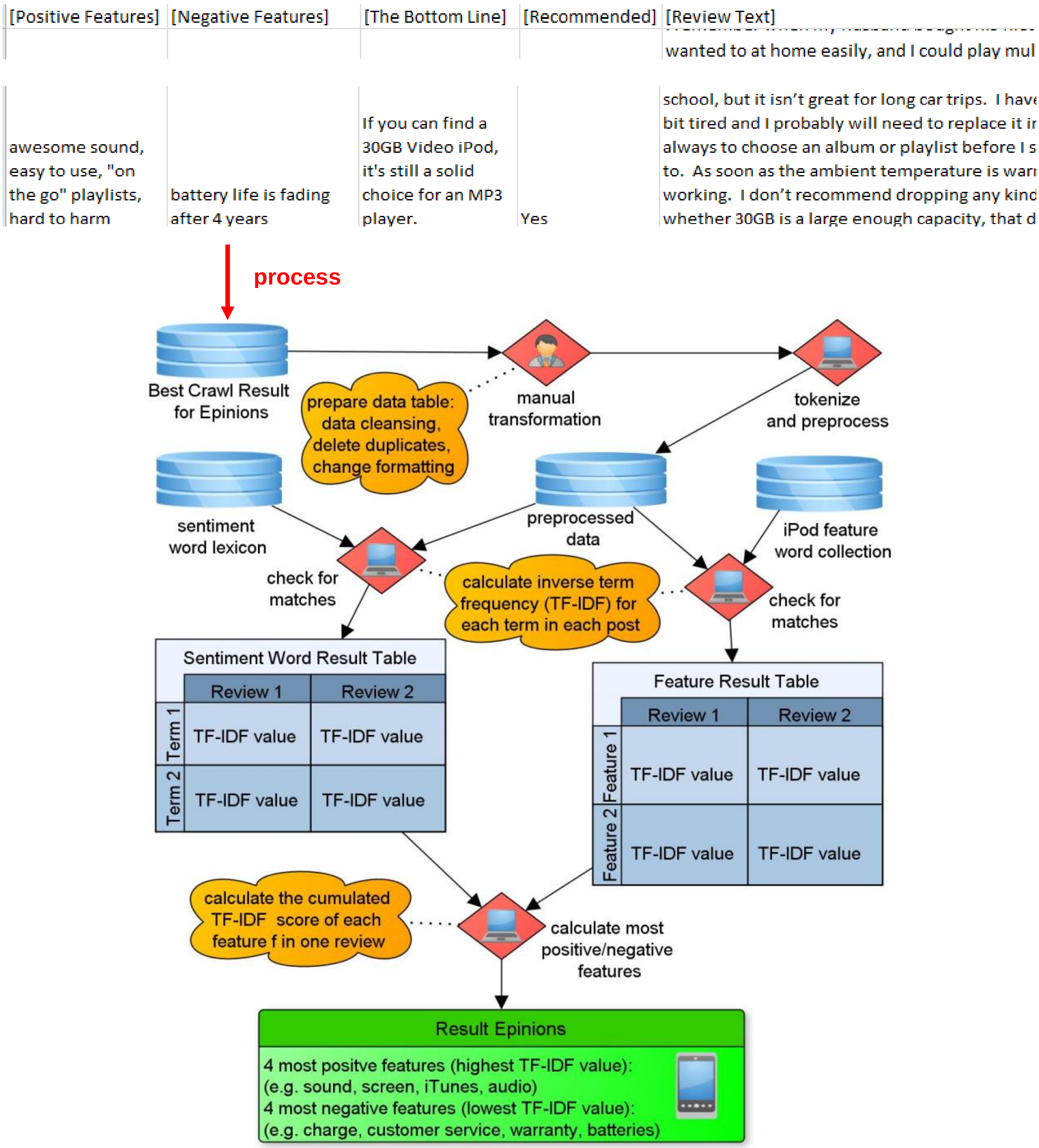
After data mining has finished, a result evaluation phase is conducted that checks if the overall sentiment classification for all reviews is correct. This is done by comparing the values like shown in the table below.

- **Overall sentiment (percentage):** Overall share of positive and negative reviews across the whole review set compared to the manually determined correct value of 86,7% and 13,3%.
- **Review sentiment correctness (percentage):** Share of correct classified separate reviews using 40 and 150 training examples.

6.1.4. Lexicon based Content Mining Epinions

The lexicon based approach allows much more detailed results. It is not limited to classifying the whole text sentiment but you can extract each single feature and its corresponding sentiment separately. To a certain extent this can be done with classification techniques as well by splitting the review text in smaller parts. For instance, a review that is structured in several chapters with different headlines covering sound, build quality, iTunes software and more can be analysed part by part but it requires manual text analysis and splitting prior to the mining process. The main reason making this impracticable is that the greater share of reviews is not written in a structured way and even if the text is separated by headlines, they most often just give a rough idea of the features the following part covers. For instance, the highest rated reviewer divides his review in the categories “Setup”, “Listening to Music”, “Watching Videos on the iPod”, “Photo Viewing” and “Other features” covering some very important but not all features the iPod has..

Using the lexicon based approach the processing order is done the way described in chapters 3.4.2 and 3.4.3. The whole mining process graph is displayed in figure 6.8. First the whole review texts are tokenized meaning they are split into single words. Afterwards preprocessing is done including converting each word to lower case. Removing stopwords[32] was not done since term filtering excluded any word not found in a lexicon anyway. Finally all reviews are divided into single sentences.



[Positive Features]	[Negative Features]	[The Bottom Line]	[Recommended]	[Review Text]
				wanted to at home easily, and I could play mul
awesome sound, easy to use, "on the go" playlists, hard to harm	battery life is fading after 4 years	If you can find a 30GB Video iPod, it's still a solid choice for an MP3 player.	Yes	school, but it isn’t great for long car trips. I have bit tired and I probably will need to replace it ir always to choose an album or playlist before I s to. As soon as the ambient temperature is wari working. I don’t recommend dropping any kind whether 30GB is a large enough capacity, that d

process
Best Crawl Result for Epinions
prepare data table: data cleansing, delete duplicates, change formatting
manual transformation
tokenize and preprocess
sentiment word lexicon
preprocessed data
iPod feature word collection
check for matches
calculate inverse term frequency (TF-IDF) for each term in each post
check for matches
Sentiment Word Result Table
Review 1
Review 2
Term 1
Term 2
TF-IDF value
Feature Result Table
Review 1
Review 2
Feature 1
Feature 2
TF-IDF value
calculate the cumulated TF-IDF score of each feature f in one review
calculate most positive/negative features
Result Epinions
4 most positve features (highest TF-IDF value):
(e.g. sound, screen, iTunes, audio)
4 most negative features (lowest TF-IDF value):
(e.g. charge, customer service, warranty, batteries)

Figure 6.8: Transformation from the crawling output table to a list of 4 most positive and 4 most negative product features using lexicon based content mining with two dictionaries.

[32] Commonly used copulas that do not contain any information like “the” or “and”

These sentences are compared with two dictionaries – one consisting of feature terms describing properties of an iPod and one consisting of positive and negative sentiment words. Each sentence containing a feature term is selected and positive and negative sentiment terms are counted. This results in a positive or negative sentiment score for the feature in the sentence. The overall score for each term was calculated by counting the sum of all TF-IDF scores for each sentence as shown in the algorithm below.

$$\boldsymbol{sentiment\ score\ for\ feature\ X} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{TFIDF\ for\ feature\ X\ in\ sentence\ i * TFIDF\ for\ sentiment\ words\ in\ sentence\ i}$$

Algorithm 6.1: Calculating the overall TF-IDF score for each sentence

To create the feature list directory, a list of iPod features as complete as possible was created by collecting features that are mentioned on Apples iPod product description page under `www.apple.com/de/ipodclassic/features.html`. After a first test round this dictionary was refined by manually analysing the test results leaving out features that did not add any value or did not address the desired features like “depth”, “height”, “lithium”, “quality” or “video”. For instance, the term “video” should address the video playback quality but was mainly used for mentioning the name “iPod video” or just stating that it had video capabilities. Further refinements were done by adding some frequently occurring variations of feature terms like “accessories”, “customer service” or “screen” that showed to be used frequently by reviewers in an evaluating way. The whole set of selected feature terms is shown in the table below.

accessories	audio	audio quality	batteries
battery	button	Buttons	capacity
charge	charging	Connection	control
controls	customer service	Display	dock
docking	earphones	Firewire	formats
frequency	hard drive	ipod with video	itunes
lcd	output	Picture	playback
power	screen	Size	sound
switch	usb	video ipod	video quality
volume	warranty	Wheel	

Table 6.1: A manually created feature list containing important properties that can be used to evaluate an iPod

The sentiment dictionary containing a list of positive and negative labelled words was taken from the Multi Perspective Question Answering website[33]. The creation of a sentiment lexicon adaptet to the domain of mp3 players containing typical sentiment words for this iPod reviews could have improved sentiment detection results. For instance, “heavy” may be a positive term when referring to the build quality of a loudspeaker but it would be clearly negative for a portable device as the iPod. The sentiment lexicon was not adapted for this thesis since the absolute result quality is not the most important factor in the tool analysis process. Instead the relative quality of each tool compared to another will be evaluated.

To prove result correctness a test set of 50 sentences is analysed looking on the features mentioned and their corresponding sentiment. The following scores are determined:

- **Precision (percentage):** How many of the feature terms found are actually what they are supposed to be? For instance in “I don't want this review sound like i hate ipods” the word

[33] `http://www.cs.pitt.edu/mpqa/subj_lexicon.html`

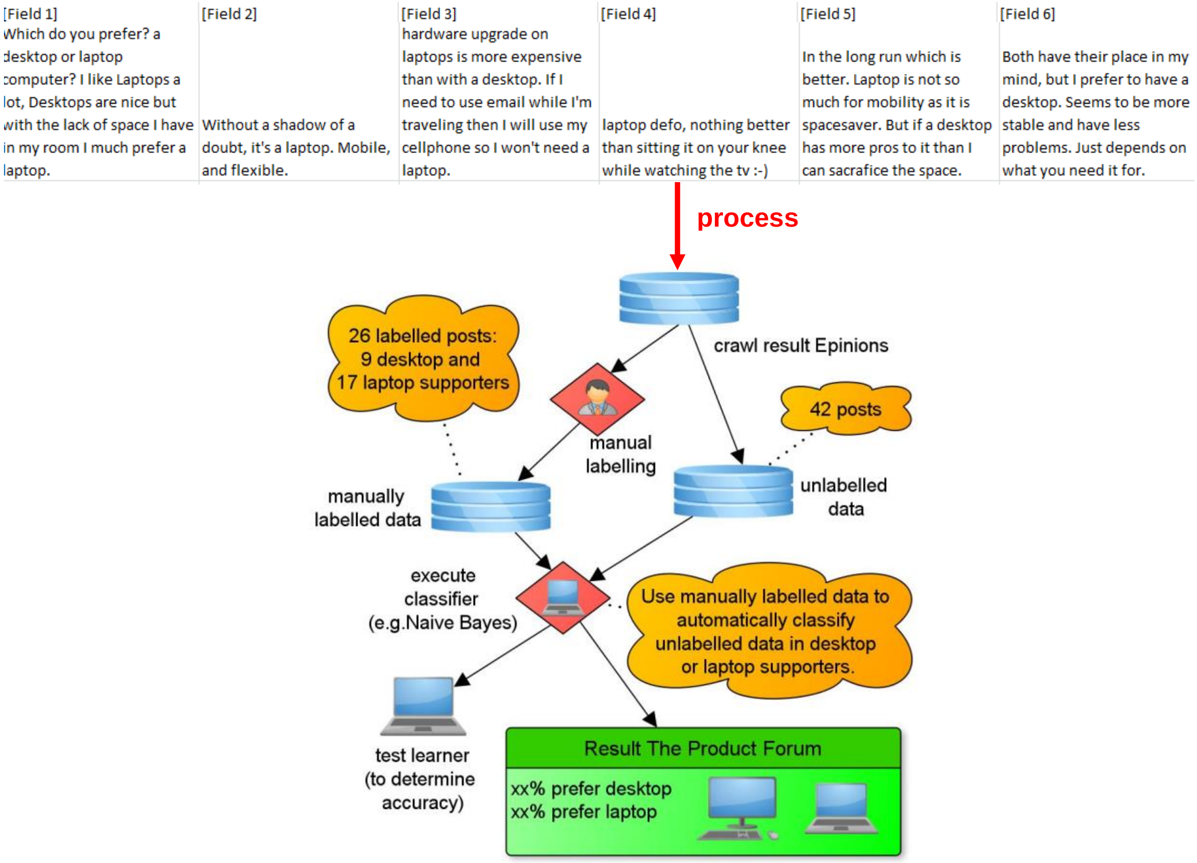
"sound" has nothing to do with the audio reproducing capability of the device and is therefore considered as failed precision.

- **Recall (percentage):** Have all features evaluated in the 50 sentences been recognized? A feature is considered as missed when it is not recognized although is clearly mentioned in a positive or negative way. Neutral statements like "the display is 3 inches tall" have no significance in evaluating the feature quality and are therefore ignored.
- **Feature sentiment correctness (percentage):** Manually vs. automatically analysed sentiment for this sentence is compared. Is the actual sentiment regarding the feature evaluated correctly?

Additionally the following test was performed on a representative random test set of up to 50 sentences including the feature under review taken from the whole set of approximately 4452 sentences that resulted after splitting the review text into single sentences counting each period as separator:

- **Feature sentiment correctness of the 4 best/worst features:** Sentiment-feature pairs are ordered by their sentiment value showing the four most positive features as well as the four most negative ones. Each feature term is checked against manual sentiment evaluation to check if the sentiment was determined correctly or not.

6.1.5. Machine Learning based Content Mining The Product Forum



[Field 1]	[Field 2]	[Field 3]	[Field 4]	[Field 5]	[Field 6]
Which do you prefer? a desktop or laptop computer? I like Laptops a lot, Desktops are nice but with the lack of space I have in my room I much prefer a laptop.	Without a shadow of a doubt, it's a laptop. Mobile, and flexible.	hardware upgrade on laptops is more expensive than with a desktop. If I need to use email while I'm traveling then I will use my cellphone so I won't need a laptop.	laptop defo, nothing better than sitting it on your knee while watching the tv :-)	In the long run which is better. Laptop is not so much for mobility as it is spacesaver. But if a desktop has more pros to it than I can sacrafice the space.	Both have their place in my mind, but I prefer to have a desktop. Seems to be more stable and have less problems. Just depends on what you need it for.

Figure 6.9: Transformation process from the crawling output table to the desired laptop and desktop supporting post percentage using supervised machine learning techniques.

The procedure for The Product Forum works very similar to the procedure for Epinions. The machine learning based classification task is conducted with a labelled test set including all 17 posts that clearly support laptops and all 9 posts favouring desktops. After being manually labelled they are used to classify all 42 posts in the thread under review to get an overall distribution of posts that prefer Desktops and Laptops.

Again classification is done using machine learning the same way like on Epinions. Opposed to Epinions an additional lexicon based classification run is performed that will be described in the next subchapter.

Looking at the device preference of forum members a manual review of all 42 postings was done resulting in the distribution table 6.2. If you do not count those posts that do not clearly prefer one device over the other, 34,6% prefer desktops and 65,4% prefer laptops.

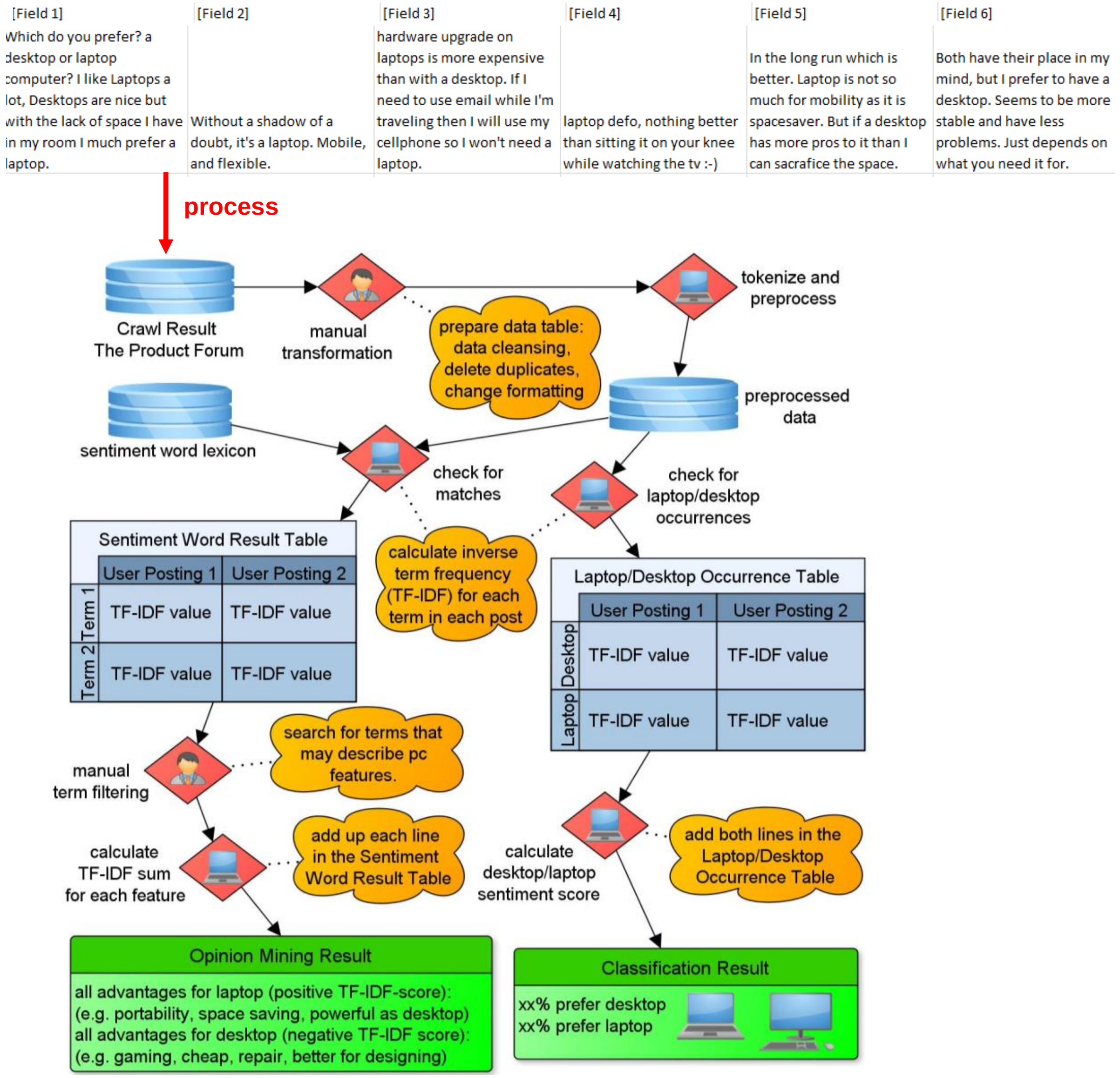
Post attitude	Number of posts	Percentage
prefer desktops	9	21,4%
prefer laptops	17	40,5%
like both equally	8	19,0%
express no attitude	8	19,0%

Table 6.2: distribution of laptop and desktop preferring posts

Evaluation is done by measuring the following value:

- **Overall preference distribution (percentage):** Share of posts classified as desktop or laptop supporters compared to reference distribution (34,6%/65,4%).

6.1.6. Lexicon based Content Mining The Product Forum



[Field 1]	[Field 2]	[Field 3]	[Field 4]	[Field 5]	[Field 6]
Which do you prefer? a desktop or laptop computer? I like Laptops a lot, Desktops are nice but with the lack of space I have in my room I much prefer a laptop.	Without a shadow of a doubt, it's a laptop. Mobile, and flexible.	hardware upgrade on laptops is more expensive than with a desktop. If I need to use email while I'm traveling then I will use my cellphone so I won't need a laptop.	laptop defo, nothing better than sitting it on your knee while watching the tv :-)	In the long run which is better. Laptop is not so much for mobility as it is spacesaver. But if a desktop has more pros to it than I can sacrafice the space.	Both have their place in my mind, but I prefer to have a desktop. Seems to be more stable and have less problems. Just depends on what you need it for.

Figure 6.10: Transformation from the crawling output table to a list of 4 positive laptop and 4 desktop features as well as to a classification of laptop and desktop supporting posts using lexicon based content mining with one feature dictionary.

Two different tasks are performed for lexicon based content mining on The Product Forum: Lexicon based classification for determining the percentage of desktop and laptop supporters and opinion mining for determining positive and negative features.

For classification the sum each of all positive and all negative sentiment words are calculated as shown in Algorithm 6.2. This would have been possible with Epinions too but it was left out since Epinions already offered this statistics on its site making this task less important. The main advantage of lexicon based classification is that no training set is required making manual labelling obsolete.

$$\boldsymbol{Desktop\ score\ (D)} = \sum_{i=1}^{n} \boldsymbol{TFIDF\ of\ positive\ Terms\ in\ sentence\ i * TFIDF\ of\ negative\ Terms\ in\ i * TFIDF\ of\ "Desktop"\ in\ i}$$

$$\boldsymbol{Laptop\ score\ (L)} = \sum_{i=1}^{n} \boldsymbol{TFIDF\ of\ positive\ Terms\ in\ sentence\ i * TFIDF\ of\ negative\ Terms\ in\ i * TFIDF\ of\ "Laptop"\ in\ i}$$

Algorithm 6.2: Calculating the overall TF-IDF score for each sentence

The sum of L and D for all posts is calculated and compared to determine which value is higher. A second calculation method is used that gives one point to the device with the higher value in a post and zero points to the lower scored device. For instance, if L is 0.5 and D is 0.2, L will get a score of 1 and D will get 0. The calculation rules are shown in Algorithm 6.3. The result that gets closer to of those two methods is taken for the comparison table.

$$\{\boldsymbol{L} > D \rightarrow L = 1; D > L \rightarrow D = 1; L = D > 0 \rightarrow L = D = 1; L = D = 0 \rightarrow L = D = 0\}$$

Algorithm 6.3: Calculation rules for absolute Laptop and Desktop mentioning values

For evaluating the lexicon based classification result the same measurement is used as for the machine learning method:

- **Overall preference distribution (percentage):** Share of posts classified as desktop or laptop supporters compared to reference distribution (34,6%/65,4%).

The lexicon based opinion mining approach is done using only one dictionary opposed to two used with Epinions. The same sentiment lexicon taken from the MPQA website is used. Since there are not a large number of features mentioned in the thread, this analysis is done without predefined features. Hence only the sentiment term scores are calculated and those terms matching to desktop or laptop features are manually chosen.

A slight difference compared to mining on Epinions website is that sentiment is calculated for a whole post and not for each single sentence separately since most postings only consist of 1-2 sentences anyway. If “laptop” and “desktop” are both mentioned in one post, this can lead to wrong conclusions. For instance, the sentence “I very much prefer a laptop over a desktop” will be counted as overall positive. Since the program cannot “understand” the semantic of the sentence both laptop and desktop are counted positive although the sentence has another meaning. The same problem occurs with features mentioned in sentences that contain both the word “desktop” and “laptop” since this method cannot judge to what device the feature belongs.

To create a reviewing base for the semi-automatically extracted feature list a manual look at the thread was taken revealing all discussed benefits or strengths of each device (see table 6.3).

Desktop features/advantages	**Laptop features/advantages**
gaming	Portability
easy to exchange parts or upgrade	space saving
cheaper repair cost	can be as powerful as desktop (except for gaming)
cheaper (on purchase)	
better for designing tasks	

Table 6.3: Manually determined desktop and laptop features

To evaluate the feature finding task the following measurement is used:

- **Number of features found (absolute number):** Every feature found during the mining process is compared to the list in table 6.3. If the feature occurs in the table one point is awarded. The more features are found the better.
- **Feature match correctness (percentage):** A check is done to determine if each feature found is matched to the correct device and the correct sentiment. The more correct matches, the better.

6.1.7. Extending the Mining Process with Stemming

Both content mining processes for Epinions and the Product Forum didn't consider one commonly used web mining step: stemming. Stemming describes the reduction of terms to their basic form. For instance, "walks", "walking", "walked" could all be reduced to the common stem "walk". This can help to reduce the number of different terms in a document and improve matching rate with lexicon terms since any word with the same stem is matched.

While case conversion and removal of stopwords like "the" or "and" are fairly straightforward and clearly enhance recall ratio as well as result clearness and tokenization is a crucial step to be able to compare the single words of a text with a lexicon, the beneficial character of stemming is not that obvious at first glance. The problem with stemming is that it implies the risk of over- and understemming. Overstemming means two words with different meanings are stemmed to the same root while understemming occurs when two words should be stemmed to a common root but are not. Besides, stemmed words do not always maintain their initial sense.

The most widely used stemming algorithms Porter and Snowball stemmer are supported by almost any tool under review. Doing some Stemming tests with KNIME using the three different stemming techniques Porter, Snowball and Kuhlen the following results could be obtained:

The review sentence:

"Once you get comfortable with itunes you realize the package is actually quite good."

Porter or Snowball stemmer turned it into:

"Onc you get comfort with itun you real the packag i actual quit good."

Kuhlen stemmer did not change anything.

Another example:

"You can add music videos and movies and t.v. all in the palm of your hand"

Kuhlen stemmer turned it into:

"You can add music and video and movy and t.v. all in the palm of your hand."

On first glance two things are getting obvious:

- Kuhlen stemmer does much less on changing terms compared to porter and the practically identical Snowball stemmer.
- A significant number of stemmed words do not make any sense.

The first finding shows that Porter or Snowball stemmer will have a much greater influence on any mining results compared to the much less aggressive Kuhlen algorithm. The second one raises the question if stemming may probably result in wrong terms leading to incorrect mining results. To prove the assumption further tests were performed using RapidMiner. In this test the mining software was used to find words in Epinions review texts that are contained in the sentiment dictionary. The number of distinct terms was used as measurement for the stemming efficiency. Measurements were done once without stemming, once using the Snowball stemmer to stem just the review text terms and once stemming all review text terms and sentiment terms as well.

	Without stemming	**Review text terms stemmed**	**Review text AND dictionary terms stemmed**
Search for negative terms	17 matches	17 matches	26 matches
Search for positive terms	46 matches	35 matches	60 matches

Table 6.4: Comparing the number of matches between the terms in the sentiment lexicon and terms in the Epinions review text corpora with and without stemming.

The result displayed in table 6.4 shows that stemming clearly enhances the term hit rate obviously caused by the generalizing nature of term stemming that decreases term variety. It still may lead to false positives as well as unrecognizable feature terms (e.g. itunes –> itun) or terms with modified sense (e.g. quite -> quit). To check the result quality, tests for Epinions and The Product Forum were performed with RapidMiner each one with and without stemming and results were compared.

	Epinions				**The Product Forum**			
	No stemming		**Stemming**		**No stemming**		**Stemming**	
	Positive	**Negative**	**Positive**	**Negative**	**Laptop**	**Desktop**	**Laptop**	**Desktop**
Cumulated TF-IDF score[34]	60,5%	39,5%	60,8%	39,2%	43,9%	56,1%	44,9%	55,1%
Absolute distribution[35]	138 92%	12 8%	143 95,3%	7 4,7%	13 59,1%	9 40,9%	18 75%	6 25%
Correct distribution	130 86,7%	20 13,3%	130 86,7%	20 13,3%	19 67,9%	9 32,1%	19 67,9%	9 32,1%
	No stemming		**Stemming**		**No stemming**		**Stemming**	
Single review/ post sentiment classification correctness	86,7%		88,7%		40,5%		45,2%	

Table 6.5: Comparing the classification accuracy of the whole review or post set as well as each single review or post with and without stemming.

What can be derived from the results presented in table 6.5? First the cumulated inverse term frequency value remains almost identical across the whole document set for both Epinions and The Product Forum with a marginal positive tendency when stemming is used. The overall sentiment distribution across all reviews too remains approximately the same regardless if the dictionaries have been stemmed or not. Epinions result got 3,3% worse while The Product Forum got 1,7% nearer to the actual result. If you evaluate the sentiment classification by counting every single review or post

[34] The sum of all sentiment term TF-IDF values in all reviews or posts

[35] The TF-IDF sentiment score for review or post is calculated separately. The absolute number of reviews having positive or negative sentiment or the number of posts preferring laptop or desktop is the result.

separately, both Epinions and The Product Forum show a little improved results when stemming is used.

This test shows an overall slight positive effect of stemming on classification especially if you classify smaller texts separately. With larger text sets results varied and got even worse for Epinions. This leads to the assumption that stemming may be advantageous for smaller text sets as they occur in opinion mining where each sentence is reviewed on its own. For classifying larger text sets the results were unclear and couldn't show significant improvements or change to the worse.

An additional test performed for The Product Forum changing the cumulated TF-IDF calculation method showed further result improvements. Every post was rated separately and the device with a higher TF-IDF value got 1 point and the lower got 0. This resulted in an increase of accuracy when using stemming for a nearly perfect results of 65,8%/34,2% compared to the real distribution of 67,9%/32,1%.

Sentiment term (stem)	**Result normal**	**Result STEMMED**	**Correct value**
+ compact	Desktop 100%	Desktop 100%	Laptop
+ light	Laptop 100%	Laptop 100%	Laptop
+ portable (portabl)	Laptop 100%	Laptop 62%	Laptop
+ stable (stabl)	Laptop 68%	Desktop 63%	Desktop
+ upgrade (upgrad)	Laptop 60%	Laptop 78%	Desktop
- break	Laptop 100%	Laptop 100%	Laptop
- expensive (expense)	Desktop 73%	Laptop 60%	Laptop
- fragile (fragil)	Desktop 60%	Laptop 59%	Laptop

Table 6.6: Percentages of post messages that assign features to laptop or desktop compared with and without the use of stemming.

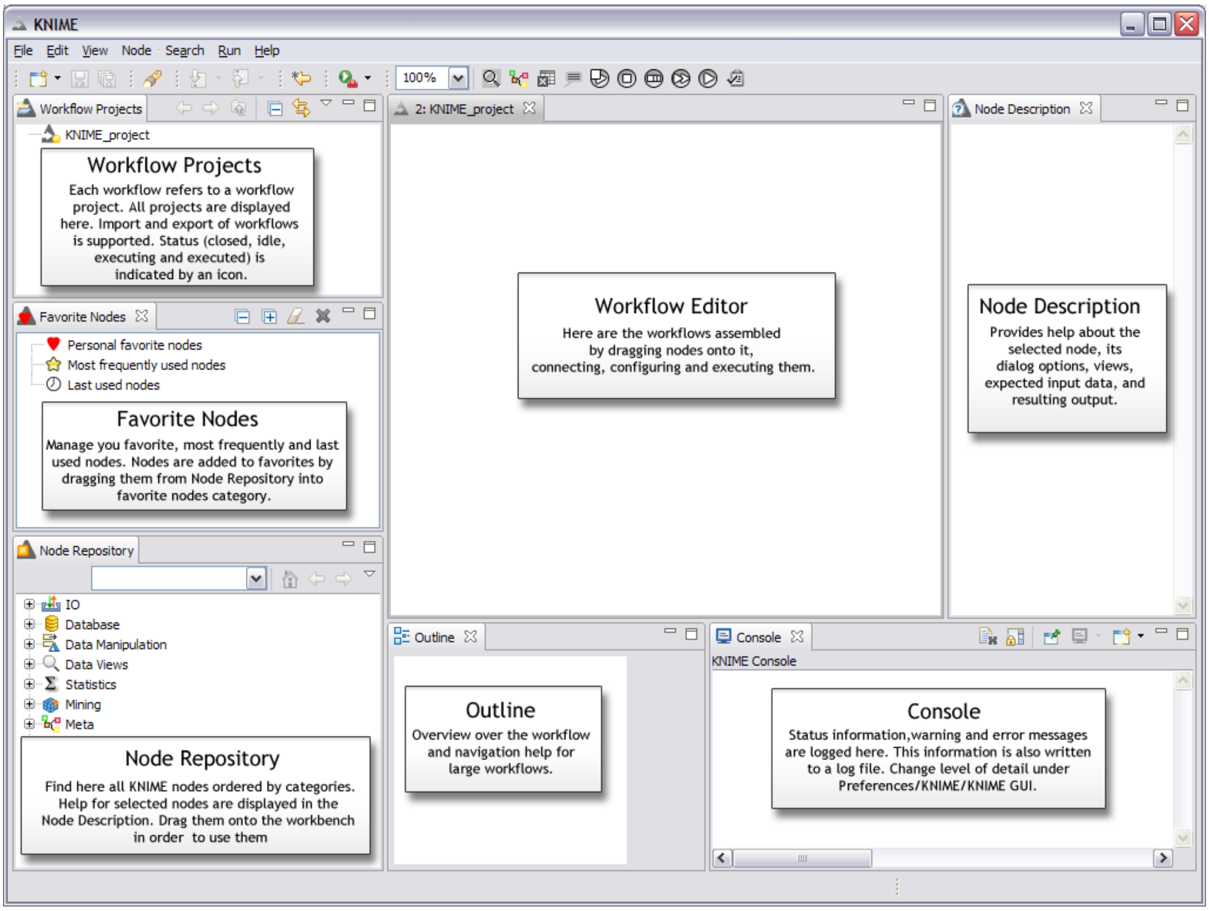
A fourth test was done shown in table 6.6 comparing sentiment word detection and classification in The Product Forum. Each percentage is calculated using the TF-IDF value of the term in a post and counting it for laptop if `TF-IDF(Laptop)>TF-IDF(Desktop)` and else for desktop. The cumulated TF-IDF of one device across all posts divided by the TF-IDF of both devices results in the relative share percentage of both devices. The device with the higher TF-IDF score is counted as related to the sentiment word. The result clearly shows a beneficial impact of stemming which increases result accuracy significantly from 3 to 6 correct values. Note that stemming increased the time necessary to select features from the sentiment word list since the higher hit rate lead to a larger term quantity and none of the added terms any feature terms.

All in all not every task could benefit from stemming and sometimes result quality even slightly decreased but overall stemming had a slightly beneficial impact on the results. Overall word hit rate could be increased as well as the correct sentiment classification of each review in Epinions and the classification of features to laptop or desktop devices for The Product Forum. The only clear result setback showed up when classifying larger text quantities like the whole set of 150 reviews from Epinions. As [50] states for a really perfect stemmer some syntactic and semantic understanding of word context would be necessary in order to avoid over-stemming or under-stemming.

Despite its slight positive impact on mining results stemming was left out from the test procedure of each tool since comparing stemmed words to original text would have made the evaluation phase more complicated since it would not suffice anymore to search features found by the software 1:1 in the source text. Instead any possible pre-stemmed form of the word would have to be considered. Besides absolute result accuracy is not that important for this thesis as relative accuracy across the tools. Just keep in mind that any results presented later could be improved a bit by using stemming.

6.2. KNIME

Installation: The Konstanz Information Miner (KNIME) is a freeware data mining tool available for Windows, Linux and Mac. To install it you just have to download the .zip archive from `http://knime.org/downloads/overview`, extract this archive and start knime.exe.



Workflow Projects

Each workflow refers to a workflow project. All projects are displayed here. Import and export of workflows is supported. Status (closed, idle, executing and executed) is indicated by an icon.

Favorite Nodes

Manage you favorite, most frequently and last used nodes. Nodes are added to favorites by dragging them from Node Repository into favorite nodes category.

Node Repository

Find here all KNIME nodes ordered by categories. Help for selected nodes are displayed in the Node Description. Drag them onto the workbench in order to use them

Workflow Editor

Here are the workflows assembled by dragging nodes onto it, connecting, configuring and executing them.

Node Description

Provides help about the selected node, its dialog options, views, expected input data, and resulting output.

Outline

Overview over the workflow and navigation help for large workflows.

Console

Status information,warning and error messages are logged here. This information is also written to a log file. Change level of detail under Preferences/KNIME/KNIME GUI.

Figure 6.11: User interface of KNIME[36]

Function principle: The core element of KNIME is the workflow editor as shown in figure 6.11. By pulling nodes from the node repository to the workflow editor and connecting them you are able to create a serial workflow. Nodes represent data processing steps and the edges connecting them represent data forwarding from one node to another. The result of one node is used as input for the next node. A possible, very simple process is shown in figure 6.12. It consists of a CSV data reader that reads the data source from a .csv file and a Table Writer that stores a KNIME data table into a file that can later be processed by other KNIME nodes.

Results can be viewed simply by right clicking the last node that should produce the desired outcome. A popup window will show the result table or graph this node has produced. This enables the user to look at intermediate results from any workflow step of the whole process.

[36] Source: `http://tech.knime.org/installation-0`

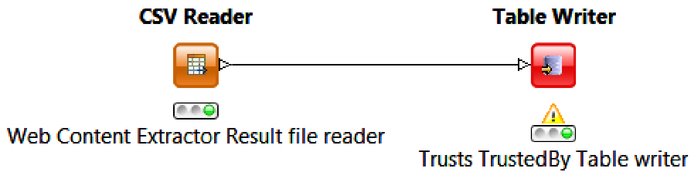

CSV Reader
Table Writer
Web Content Extractor Result file reader
Trusts TrustedBy Table writer

Figure 6.12: A simple workflow in KNIME: Reading data from a .csv file and forwarding it to the second node that writes a KNIME data table.

In Order to enhance the functionality of KNIME there exist a multitude of extensions each consisting of up to 40 additional nodes. Each extension can be installed separately simply by selecting the appropriate menu point in the software menu. Following KNIME Labs Extensions have been added in to enable the software to execute the desired structure and content mining tasks:

a) Network Mining
b) Textprocessing

Two other extensions – “Webanalytics” and “Indexing and Searching” were tested as well but they did not add any value to this evaluation task. While “Indexing and Searching” only allows searching the web or websites for specific search terms making it inappropriate for the crawling task in online communities, “Webanalytics” offers mainly methods to get rid of html code and other noise that was not existent in the results from Web Content Extractor used as input for KNIME. However, it may be helpful for preprocessing noisy crawling results as produced with Winweb Crawler 2.0.

Introduction and examples for how you can perform structure and text mining as desired is given in a number of papers [51][52][53][54].

6.2.1. Structure Mining Epinions

With the installed network mining plugin we are able to display the user A trusts user B relationships as a directed graph similar to the reference graph.

First the result table from Web Content Extractor software had to be converted into the format required by KNIME consisting of two columns containing the trust expressing persons edge in column 1 and the trust receiving edge in column 2. Then the “Network Plugin” node could be used to create a directed graph that correctly visualized the user-trusts-user relationships.

Data import turned out to be rather complicated since the built-in CSV reader would not read cell values if they contain line breaks. Prior to importing the CSV file manual preprocessing in a text editor was necessary to delete disturbing line breaks. This can be a rather time consuming procedure for larger datasets.

After processing the collected data and creating a network graph the result was displayed correctly like in the reference graph. Unfortunately the amount of information displayed in the unfiltered graph shown in figure 6.13 is overwhelming as already discussed in chapter 6.1.1. This is because a person B expressing his trust to another person A can have other persons C, D, E expressing their trust to B making it necessary to evaluate persons C, D and E as well and so on, making the graph huge and unclear.

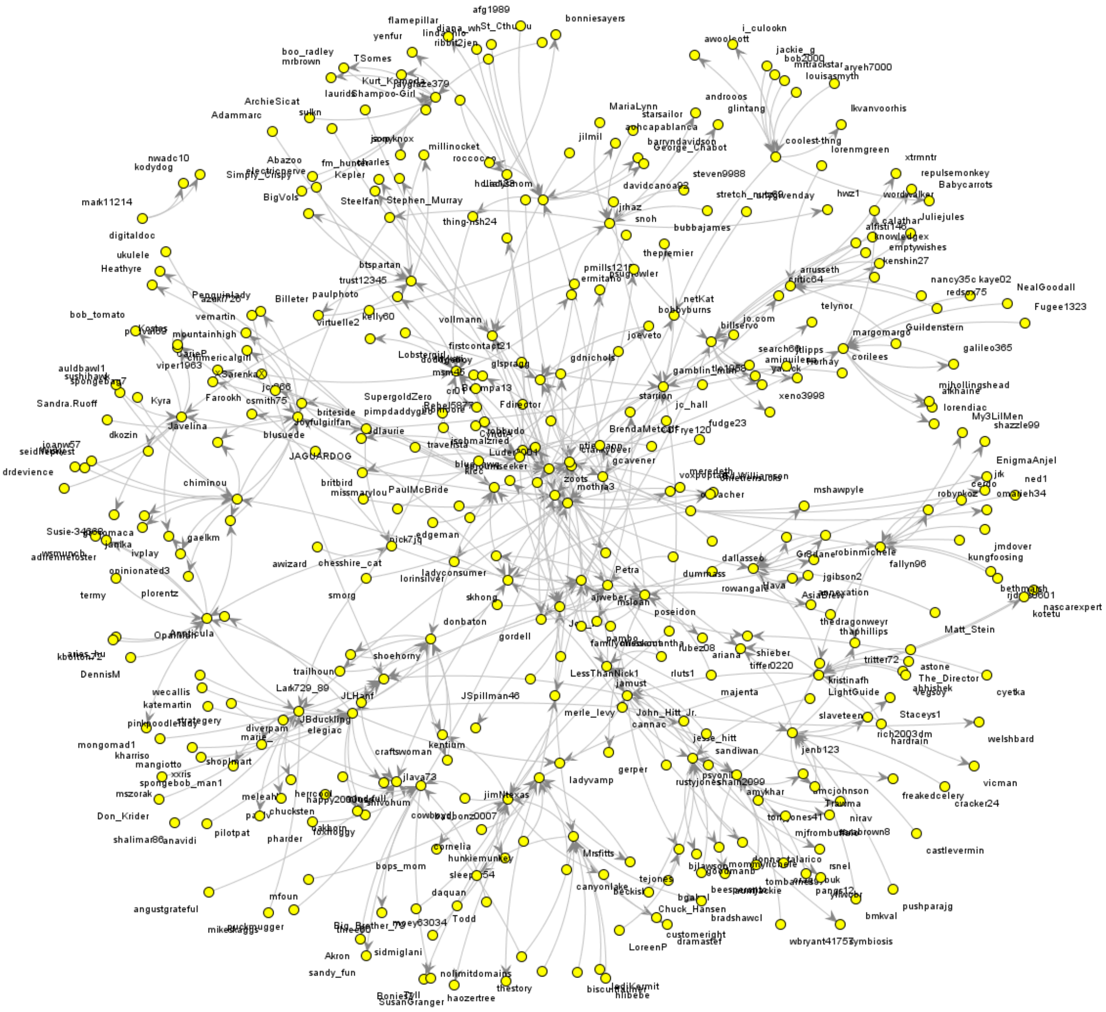


Figure 6.13: The result graph of Epinions including all users

To make results more clearly arranged KNIME offers the possibility to extract a subgraph containing only those nodes that are nearly related to the tree users under review (shieber, joeveto, nwadc10). This means only those persons expressing their trust directly to one of those three or those expressing their trust to a person that directly expresses their trust to one of the tree are taken into account. The resulting subgraph is much more clearly arranged and is shown in figure 6.14.

The result comes very near to the manually created reference graph if you only consider those users very closely related to the three core reviewers that are marked red in the reference graph. The only small drawback compared to the reference was the inability of KNIME for clearer automatic edge alignment and a missing option for manual edge position adjustment. The informations that can be derived from the subgraph are the same as from the reference graph. Names and connections are all there to their full extent. No errors were detected.

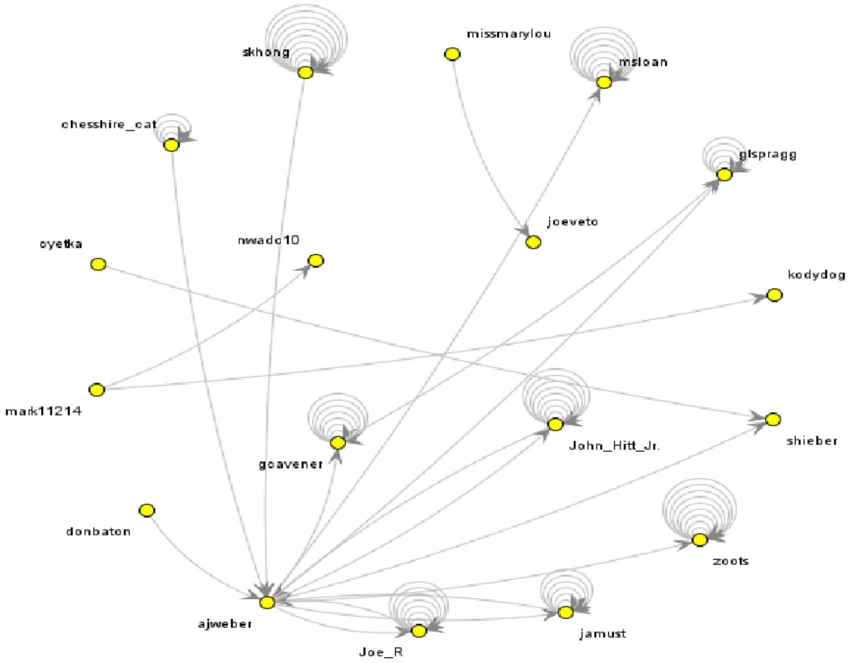


Figure 6.14: A subgraph within a certain range of three users (shieber, joeveto, nwadc10)

6.2.2. Structure Mining The Product Forum

Structure mining The Product Forum was conducted in a similar way as Epinions except for an additionally filter workflow was used with a “Dictionary Tagger” node also used for used for sentiment word filtering in the lexicon based content mining tasks. The workflow was programmed to search every post message for author names to generate a list of author names in the left column and member names the author mentioned in the right column.

The result graph in figure 6.15 shows relationships between forum member posts in a correct way as it displays identical information as the reference graph.

It is possible to display post messages as edge labels in the graph but tests showed that this display method is very confusing. Messages can be displayed in one line but they cannot be truncated like in the reference graph as well resulting in long word chains that go well beyond the edge lengths ending at the graph window boundaries making the graph almost unreadable.

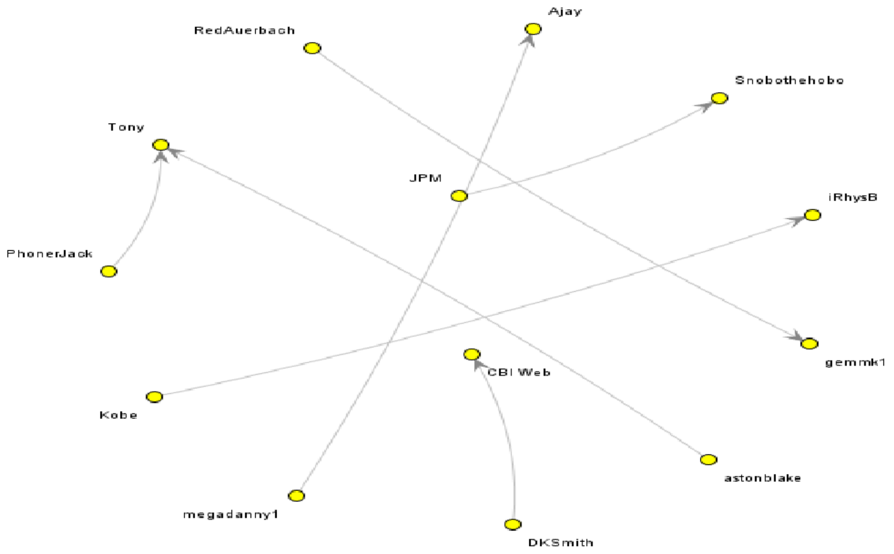


Figure 6.15: Structure graph showing which user mentions another user in his post message.

6.2.3. Machine Learning based Content Mining Epinions

KNIME is capable of content mining with a lexicon based method as well as text classification based.

KNIME supports training of models with its “cross validation” nodes but unfortunately they are not intended for use on unlabelled data. Instead it is stated in KNIME’s forum that a model can only be learned with a given test set and tested on its efficiency with the same test set. Hence, classifying all unlabelled reviews by using a small set of manually labelled data is not possible.

Note that Naive Bayes and Support Vector Machine classifier did not lead to any different results hence no further test comparing these two methods were performed.

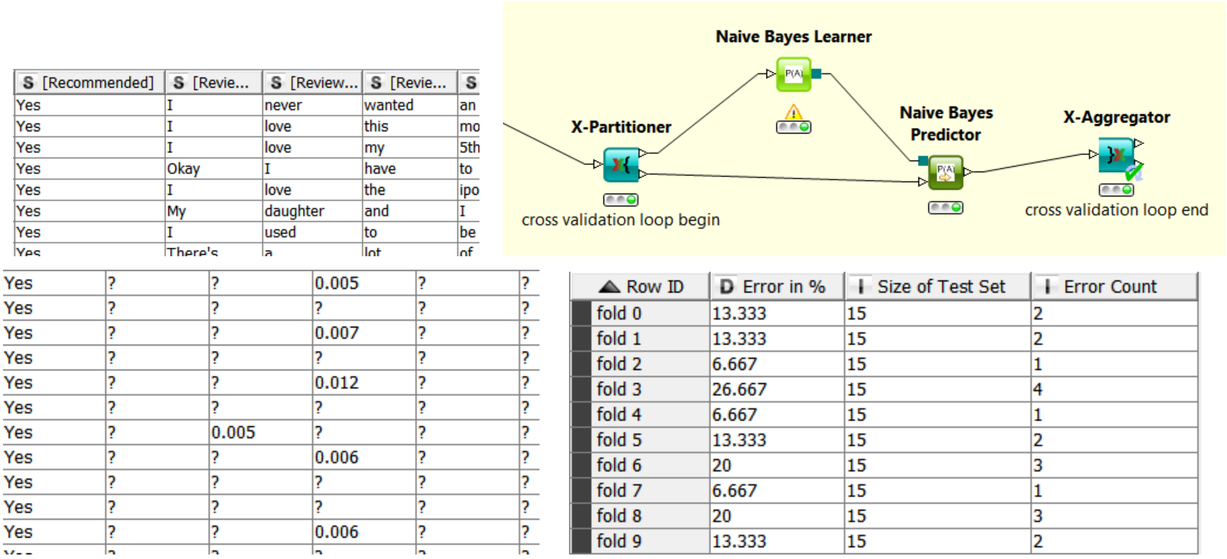
Two different tokenizing methods supported by KNIME were used for the classification test. The first was the most simple by breaking down the review text into single words using white space as separation method but with this method not the words itself are compared during the learning step but the word positions. The second one used the built-in relative term frequency calculation in a pivot table. The relative term frequency is calculated as shown in algorithm 6.4.

$$\boldsymbol{relative\ term\ frequency\ for\ term\ t} = \frac{\boldsymbol{occurrences\ of\ t\ in\ this\ document}}{\boldsymbol{occurrences\ of\ t\ in\ all\ documents}}$$

Algorithm 6.4: Calculating the overall TF-IDF score for each sentence

Unfortunately, although the result accuracy as displayed in figure 6.16 on the bottom right seems to be decent at first glance with only 21 out of 150 reviews classified wrong, looking at the results in deeper detail revealed that all 150 reviews were classified as positive using the single word splitting tokenization method and 149 were classified as negative using the relative term frequency method. Unfortunately even the 1 single negative review was classified wrong. This means that, despite offering high result accuracy in this case, accuracy in other, more even distributed data sets is more than questionable. To test performance on even distributed data the 40 example set consisting of even distributed 20 positive and 20 negative reviews were used and classified using the relative term frequency preparation method. It showed result correctness of only 15 out of 40 reviews with overall

12 positive and 28 negative classified reviews. A random classifier would have been the better choice in this case.



S [Recommended]	S [Revie...	S [Review...	S [Revie...	S
Yes	I	never	wanted	an
Yes	I	love	this	mo
Yes	I	love	my	5th
Yes	Okay	I	have	to
Yes	I	love	the	ipo
Yes	My	daughter	and	I
Yes	I	used	to	be
Yes	There's	a	lot	of

Yes	?	?	0.005	?	?
Yes	?	?	?	?	?
Yes	?	?	0.007	?	?
Yes	?	?	?	?	?
Yes	?	?	0.012	?	?
Yes	?	?	?	?	?
Yes	?	0.005	?	?	?
Yes	?	?	0.006	?	?
Yes	?	?	?	?	?
Yes	?	?	?	?	?
Yes	?	?	0.006	?	?
Yes	?	?	?	?	?

Row ID	D Error in %	I Size of Test Set	I Error Count
fold 0	13.333	15	2
fold 1	13.333	15	2
fold 2	6.667	15	1
fold 3	26.667	15	4
fold 4	6.667	15	1
fold 5	13.333	15	2
fold 6	20	15	3
fold 7	6.667	15	1
fold 8	20	15	3
fold 9	13.333	15	2

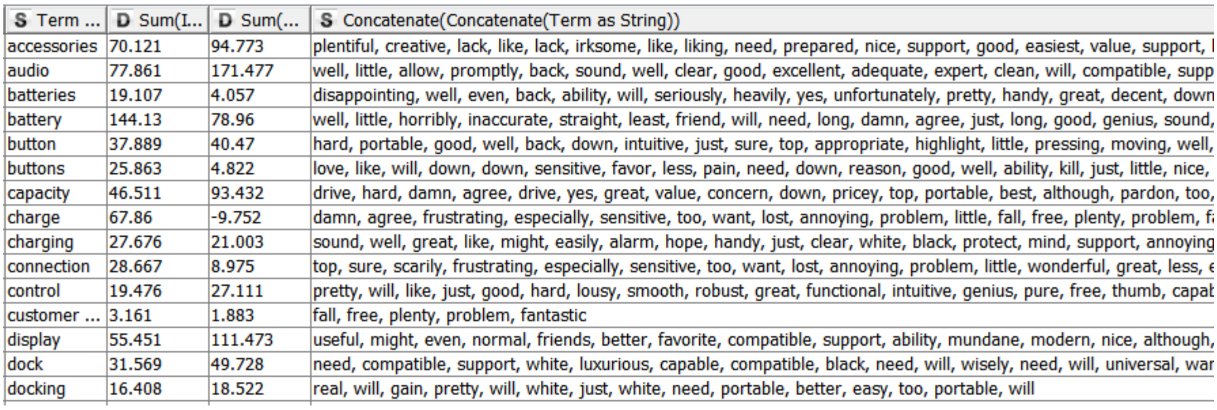
Figure 6.16: Left: Two different input formats (text breakdown in words on top and term frequency pivot table on the bottom). Right: Cross validation process on the top and result accuracy on the bottom.

	Single word breakdown	Relative term frequency	Correct value
Overall sentiment positive/negative 40	na	na	86.7%/13,3%
Overall sentiment positive/negative 150	100%/0%	99,3%/0,7%	86.7%/13,3%
Review sentiment correctness 40	na	na	100%
Review sentiment correctness 150	86,7%	86%	100%

Table 6.7: Results evaluated with both preprocessing methods. The overall sentiment was determined much too positive biased. Since classification did not work for unlabelled data, no results could be obtained when using 40 training examples.

6.2.4. Lexicon based Content Mining Epinions

After preprocessing was done review sentences were compared both dictionaries and those words that matched dictionary terms were tagged with KNIMEs Standard Named Entity Filter Node. After several workflow steps including the use of self-written Java snippets to split, extract, combine and filter data as well as calculate the desired TF-IDF values the result table looked like figure 6.17.



S Term ...	D Sum(I...	D Sum(...	S Concatenate(Concatenate(Term as String))
accessories	70.121	94.773	plentiful, creative, lack, like, lack, irksome, like, liking, need, prepared, nice, support, good, easiest, value, support, l
audio	77.861	171.477	well, little, allow, promptly, back, sound, well, clear, good, excellent, adequate, expert, clean, will, compatible, supp
batteries	19.107	4.057	disappointing, well, even, back, ability, will, seriously, heavily, yes, unfortunately, pretty, handy, great, decent, down
battery	144.13	78.96	well, little, horribly, inaccurate, straight, least, friend, will, need, long, damn, agree, just, long, good, genius, sound,
button	37.889	40.47	hard, portable, good, well, back, down, intuitive, just, sure, top, appropriate, highlight, little, pressing, moving, well,
buttons	25.863	4.822	love, like, will, down, down, sensitive, favor, less, pain, need, down, reason, good, well, ability, kill, just, little, nice,
capacity	46.511	93.432	drive, hard, damn, agree, drive, yes, great, value, concern, down, pricey, top, portable, best, although, pardon, too,
charge	67.86	-9.752	damn, agree, frustrating, especially, sensitive, too, want, lost, annoying, problem, little, fall, free, plenty, problem, fa
charging	27.676	21.003	sound, well, great, like, might, easily, alarm, hope, handy, just, clear, white, black, protect, mind, support, annoying
connection	28.667	8.975	top, sure, scarily, frustrating, especially, sensitive, too, want, lost, annoying, problem, little, wonderful, great, less, e
control	19.476	27.111	pretty, will, like, just, good, hard, lousy, smooth, robust, great, functional, intuitive, genius, pure, free, thumb, capab
customer ...	3.161	1.883	fall, free, plenty, problem, fantastic
display	55.451	111.473	useful, might, even, normal, friends, better, favorite, compatible, support, ability, mundane, modern, nice, although,
dock	31.569	49.728	need, compatible, support, white, luxurious, capable, compatible, black, need, will, wisely, need, will, universal, war
docking	16.408	18.522	real, will, gain, pretty, will, white, just, white, need, portable, better, easy, too, portable, will

Figure 6.17: Content Mining result table for Epinions showing features in the first column followed by the sum of TF-IDF-occurrence values in column 2, sentiment values in column 3 and all words co-occurring with the feature term in column 4.

Figure 6.18 shows a graphical display supported by KNIME, a conditional box plot of the cumulated TF-IDF values. It shows that three features are outstanding among all the others getting sentiment values more than twice as high: sound, screen and iTunes. Since only one feature gets a slightly negative value this leads to the assumption that reviews were formulated in an overall rather positive way. Since the overall score on Epinions for the iPod is greater than 4 out of 5 stars this assumption makes sense.

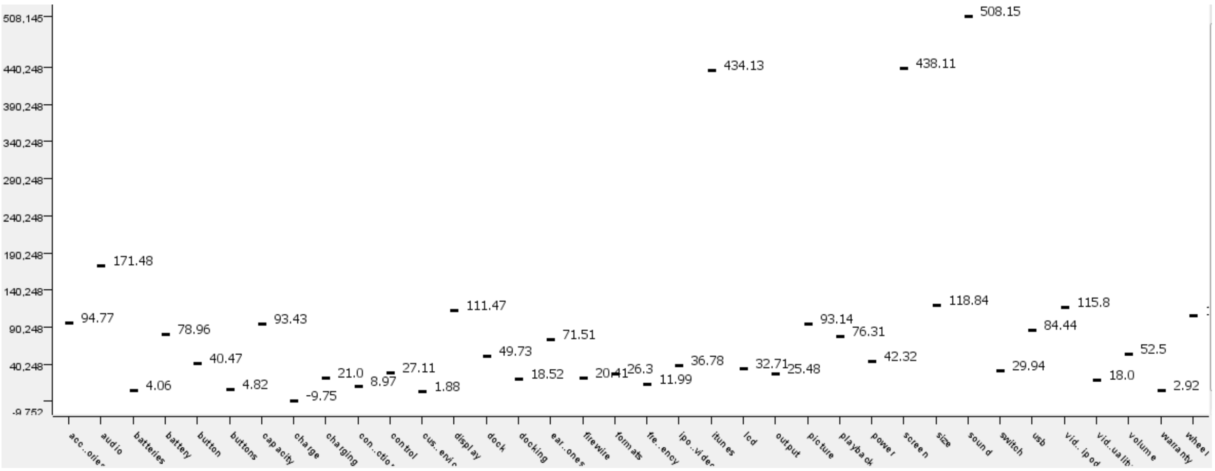


Figure 6.18: Conditional box plot comparing the TF-IDF sentiment values with each other

Another graphical representation created by KNIME is a pie chart shown in figure 6.19. It compares the number of mentions of every term showing that most reviewers talk about iTunes, the battery (runtime), screen or sound. A third of all mentions account to these four terms although 36 feature terms were mentioned altogether.

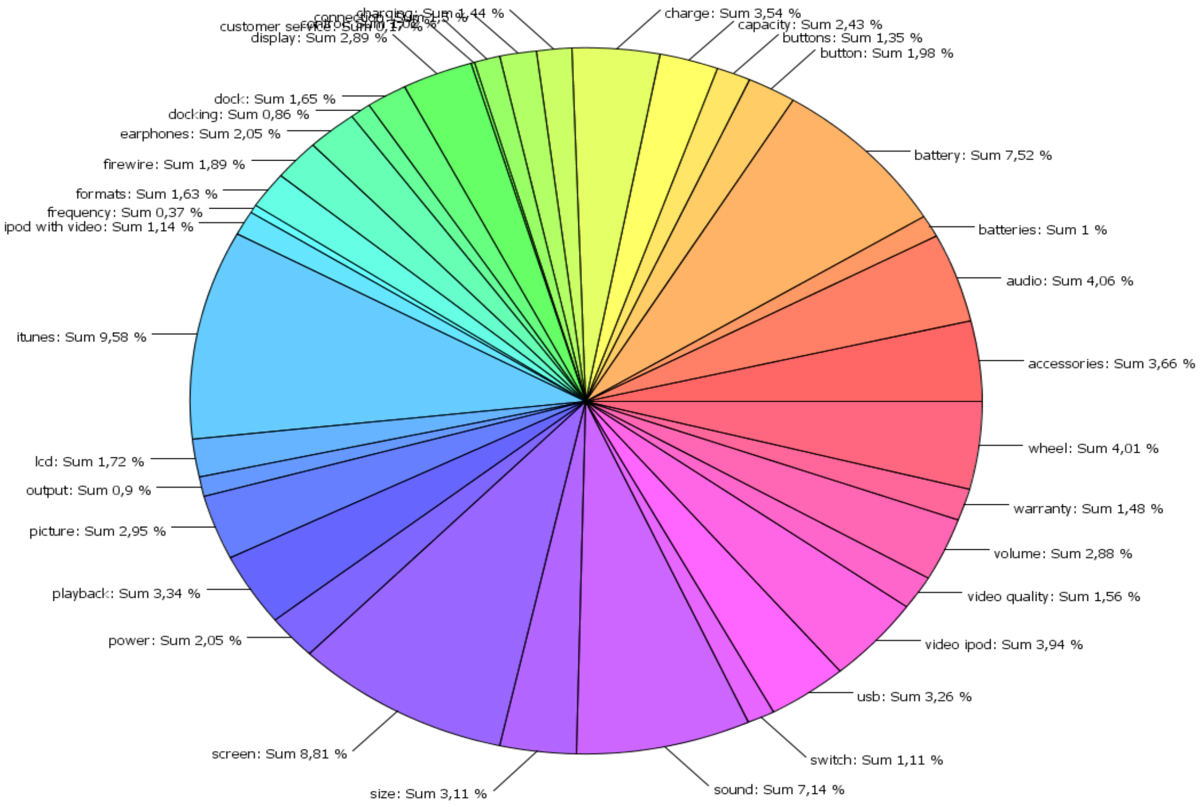


Figure 6.19: Comparing the cumulated share of IDF-scores of each term showing which terms were discussed most frequently (iTunes, screen, battery and sound)

Features found	TF-IDF value by KNIME	Real sentiment
sound (positive)	508,15	positive (27 positive, 21 neutral, 2 negative mentions in 1585 sentences)
screen (positive)	438,11	positive (18 positive, 27 neutral, 5 negative mentions in 915 sentences)
iTunes (positive)	434,13	positive (12 positive, 31 neutral, 7 negative mentions in 464 sentences)
audio (positive)	171,48	positive (7 positive, 38 neutral, 5 negative mentions in 2843 sentences)
charge (negative)	-9,75	negative (2 positive, 33 neutral, 15 negative mentions in 1494 sentences)
customer service (negative)	1,88	positive (1 positive, 0 neutral, 0 negative mentions in 4452 sentences)
warranty (negative)	2,92	negative (1 positive, 15 neutral, 3 negative mentions in 4452 sentences)
batteries (negative)	4,06	negative (0 positive, 6 neutral, 4 negative mentions in 4452 sentences)

Table 6.8: Feature finding result showing the 4 most positive and the 4 most negative rated features compared to their manual evaluated real features.

Table 6.8 shows the feature sentiment correctness of each 4 most positive and most negative rated features. Since for all positive and 3 out of 4 negative classified terms the sentiment could be confirmed during manual review, the results are quite satisfying.

The first positive classified term “sound” occurred 27 times in a positive way, 21 times it was mentioned in a neutral way or off topic (e.g. “I do not want this review to sound like…” has nothing to do with how the device playback sounds) and only 2 times in a negative way. Evaluation was done to the 50th occurrence of the feature term hence manual evaluation stopped at sentence 1585 out of a total of 4452 sentences. The positive sentiment can be confirmed for this term.

The other positive classified terms turned out to be true positives as well and surprisingly even the order of terms was determined correctly with the positive/negative ratio getting closer with the 4th term “audio” having only a slight positive tendency. Audio had the greatest share of neutral and off-topic mentions with only 24% really speaking about audio quality of the device. Off topic mentions were numerous and talked mainly about audio plugs or audio formats the device supports.

Looking at the negative features one major difference is visible at the first glance: The number of term occurrences is overall way below the positive terms. This raises the question if simply adding up the TF-IDF scores of all sentences may favour terms occurring more often and if it lets infrequent terms appear worse than they are.

Still 3 out of 4 negative classified terms showed to be true negatives. The most negative classified term “charge” had 15 negative opposing to only 2 positive mentions and most reviewers complained about the missing charger in the retail package as well as the short lived battery that survives only a few hundred charges. The same opinions showed up on the feature “batteries” where reviewers complained about short battery runtime and the incapability to replace the battery.

The feature that could not be determined correctly was “customer service” that had only one single occurrence throughout the whole review set. This apparently raises the chance for randomly occurring classification errors. Maybe the introduction of a term occurrence filter leaving out those features that occur less than about 3-5 times could decrease the probability for such errors in future experiments.

	Absolute value	**Percentage**
Precision	104/121	85,95%
Recall	106/110	96,36%
Sentiment correctness	22 correct, 23 neutral, 5 wrong	81,48%

Table 6.9: Evaluation of three metrics precision, recall and sentiment correctness using a test set of 50 sentences

Table 6.9 shows the other three metrics evaluated as defined in chapter 6.1.4. While precision and recall are rather good meaning the features itself are identified correctly most of the times, the sentiment correctness is not that clearly trustworthy. While the share of correctly determined sentiment sentences is perfectly fine, a large number of sentences were found that either do not express any sentiment (mentioned as neutral in the table) or positive and negative sentiments as well. For instance, “thanks to itunes you no longer have to separately download clumsy software updates but I’ve never bought into the ipod locks you into itunes gripe” mentions itunes in a positive and a negative way. If the sentiment analysis counts it as positive, this is correct, but incomplete since it should count it as negative as well. Counting such sentences as either positive or negative may result in a biased sentiment evaluation decreasing sentiment detection accuracy.

6.2.5. Machine Learning based Content Mining The Product Forum

Like with Epinions classification was done using either the preparation step of dividing texts into single words or creating a term frequency matrix just as shown in chapter 6.2.3, figure 6.15, left. The results looked similar to Epinions. Again nearly all text fragments were classified in the dominating class as shown in table 6.10.

	Single word breakdown	Term frequency	Correct value
Overall distribution laptop/desktop 26	92,3%/7,7%	100%/0%	65,4%/34,6%
Classification correctness 26	73,1%	65,4%	100%
Classification correctness 18	66,7%	22,2%	100%

Table 6.10: Classification result using 26 or 18 posts as training and classification sets.

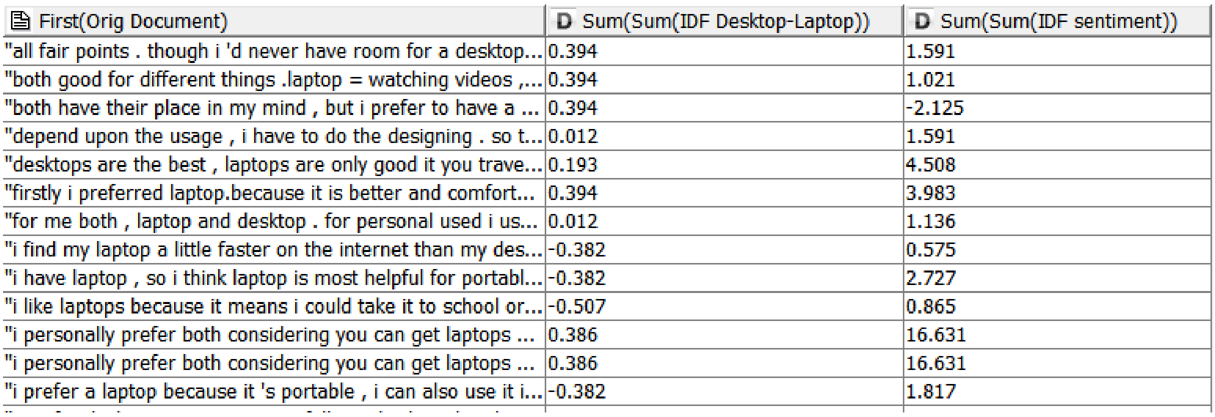
Accuracy seems quite good when looking at the classification correctness numbers but the same drop in accuracy arises as with Epinions when changing the training set to an even distributed set containing 9 positives and 9 negatives. Using the term frequency matrix as source only 4 out of 18 (22,22%) could be classified correctly while the word breakdown method still worked quite good with 12 out of 18 (66,67%) correct classified values. Overall, the word breakdown method gives comparatively decent results at least when classifying shorter texts.

6.2.6. Lexicon based Content Mining The Product Forum

Performing classification on lexical basis was done as described in chapter 6.1.6 using two different distribution calculation methods. While calculation with algorithm 2 failed to deliver correct results, using algorithm 3 delivered sufficiently accurate results showing overall correct tendency favouring laptops.

	Laptop	Desktop
Overall preference distribution (algorithm 2 – TF-IDF sums)	31,9%	68,1%
Overall preference distribution (algorithm 3 – absolute preference count)	53,6%	46,4%
Correct distribution value	65,4%	34,6%

Table 6.11: Evaluation of post classification using the two calculation methods presented in chapter 6.1.6



First(Orig Document)	Sum(Sum(IDF Desktop-Laptop))	Sum(Sum(IDF sentiment))
"all fair points . though i 'd never have room for a desktop...	0.394	1.591
"both good for different things .laptop = watching videos ,...	0.394	1.021
"both have their place in my mind , but i prefer to have a ...	0.394	-2.125
"depend upon the usage , i have to do the designing . so t...	0.012	1.591
"desktops are the best , laptops are only good it you trave...	0.193	4.508
"firstly i preferred laptop.because it is better and comfort...	0.394	3.983
"for me both , laptop and desktop . for personal used i us...	0.012	1.136
"i find my laptop a little faster on the internet than my des...	-0.382	0.575
"i have laptop , so i think laptop is most helpful for portabl...	-0.382	2.727
"i like laptops because it means i could take it to school or...	-0.507	0.865
"i personally prefer both considering you can get laptops ...	0.386	16.631
"i personally prefer both considering you can get laptops ...	0.386	16.631
"i prefer a laptop because it 's portable , i can also use it i...	-0.382	1.817

Figure 6.20: Part of KNIME lexicon based classification result table as output

Some strange program behaviour could be observed during testing was that the "GroupBy" node sometimes seemed to deliver false results. It should group identical feature term lines and calculates the sum of the corresponding term frequencies. The calculation was done correctly but it left some duplicate lines in the result table during several steps in the process as can be seen in figure 6.20. Evaluation of preference share distribution was done ignoring those duplicates and counting only distinct lines.

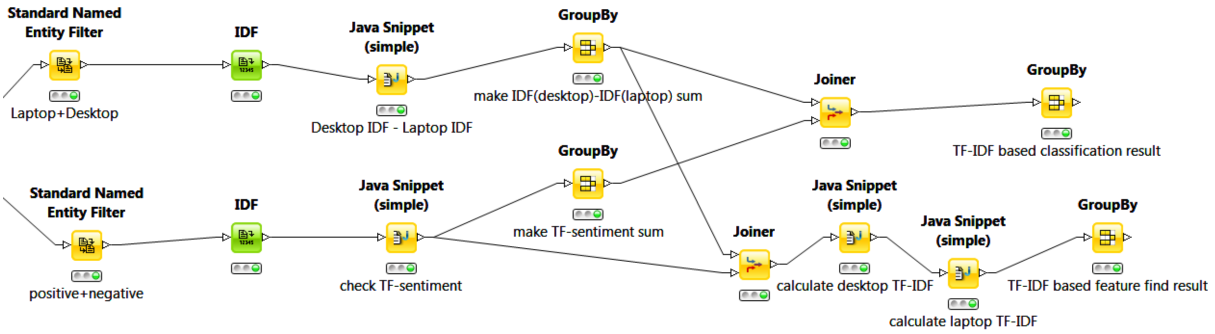


Figure 6.21: Term frequency calculation, classificaton and feature extraction process modelled in KNIME after the lexica have been preprocessed and the data source have already been tagged with the "Dictionary Tagger" node.

Extracting words from the sentiment lexicon and comparing them with the features found during manual evaluation of the thread revealed a feature list as shown in table 6.12. Notice that feature unspecific words like “better”, “prefer”, “will”, “easily”, “need”, “less”, etc. were excluded from the list because they contain no feature information. This exclusion could only be done in a manual way.

Manual review		**Supervised automatic review**			
Laptop	**Desktop**	**Laptop pos.**	**Laptop neg.**	**Desktop pos.**	**Desktop neg.**
Portability	Gaming	flexible (1,59)	expensive (4,85/2,42)	compatible (3,41/2,27)	break (1,59)
space saving	easy to exchange parts or upgrade	portable (3,41)	fragile (2,60)	compact (2,60/2,60)	cheap (1,59)
can be as powerful as a desktop (except for gaming)	cheaper repair cost	stable (1,30/1,30)		stable (1,30/1,30)	
	Cheaper	compact (2,60/2,60)		comfortable (2,28)	
	better for designing tasks	upgrade (5,32/2,28)			

Table 6.12: Manual vs. automated feature recognition: dark green fields are classified correctly, orange fields are matched to the false device or sentiment. Light green are features that make sense in the context but do not match actual existing features. The values in brackets show the TF-IDF compared to the TF-IDF given to the other device (if any).

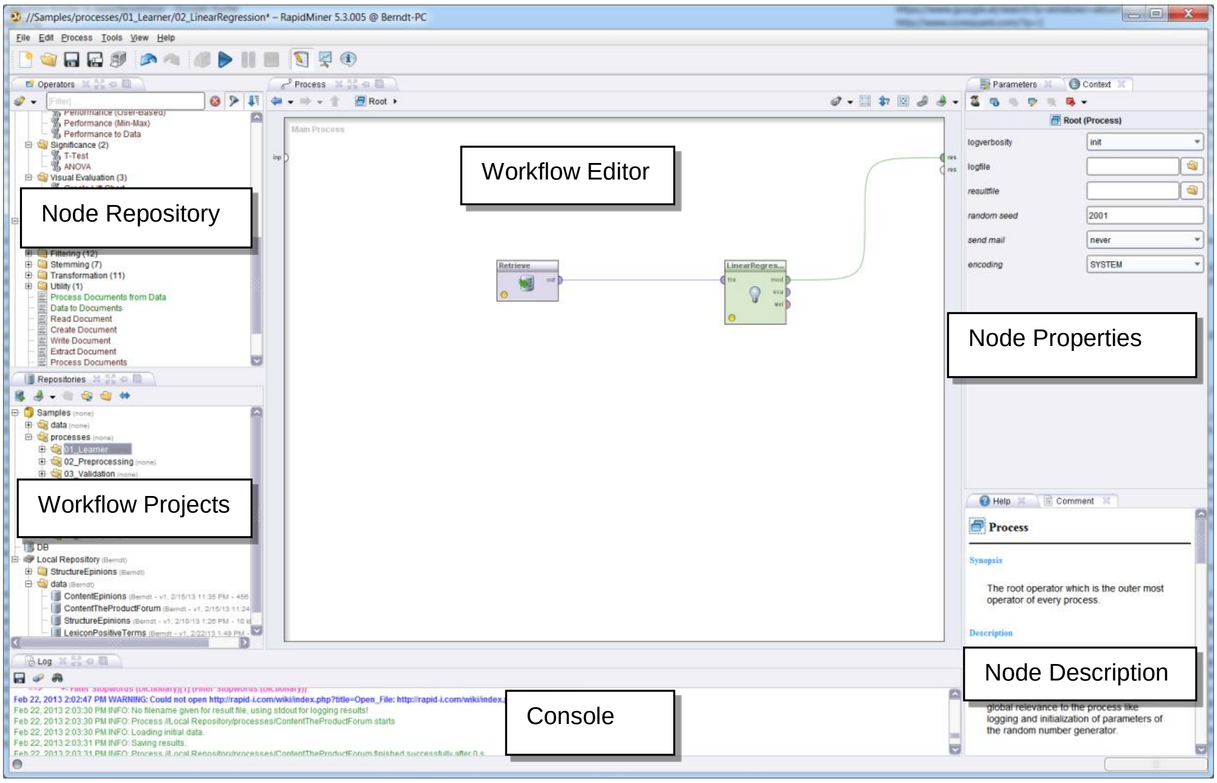
As you can see 4 out of 8 features recognized during the manual review could be determined, namely the features “portable”, “compact”(=“space saving”), “upgrade” and “cheap”(= opposite of “expensive”). The assignment to the device type and sentiment did not work too well since exactly 50% were assigned wrong making the result seem pretty random. In contrast to the evaluation for Epinions which used a manually created feature list, this approach worked without such a list resulting in clearly lower accuracy.

	Absolute number	**Percentage**
Number of features found	4	50%
Feature match correctness (correct device AND sentiment)	3	50%

Table 6.13: Evaluation metrics for features found

6.3. Rapid Miner

Installation: The Rapid Miner executable install file can be downloaded from the Rapid-I homepage by following the URL `http://rapid-i.com/content/view/181/190/`. Executing starts the automatic installation routine under Windows.



logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM

The root operator which is the outer most operator of every process.

global relevance to the process like logging and initialization of parameters of the random number generator.

```
Feb 22, 2013 2:02:47 PM WARNING: Could not open http://rapid-i.com/wiki/index.php?title=Open_File: http://rapid-i.com/wiki/index.
Feb 22, 2013 2:03:30 PM INFO: No filename given for result file, using stdout for logging results!
Feb 22, 2013 2:03:30 PM INFO: Process //Local Repository/processes/ContentTheProductForum starts
Feb 22, 2013 2:03:30 PM INFO: Loading initial data.
Feb 22, 2013 2:03:31 PM INFO: Saving results.
Feb 22, 2013 2:03:31 PM INFO: Process //Local Repository/processes/ContentTheProductForum finished successfully after 0 s
```

Figure 6.22: User inferface of Rapid Miner

Function Principle: Similar to KNIME Rapid Miner uses nodes connected through edges to represent workflows. An edge forwards the result data from the source node to the input edge on the target node. The most noticeable difference in the user interface is that it consists of two data views – the design and the result view. In KNIME you could view result tables or graphs directly from a selected node in the workflow editor by opening a popup. Rapid Miner uses another approach showing the results in a separate view. Otherwise the two programs work similar which means you can easily learn how to use the other program if you already know how to use one of them.

As with KNIME the standard node repository can be extended from within the software menu. For web mining tasks the following extensions had to be installed from the "Help" –> "Updates and Extensions" menu:

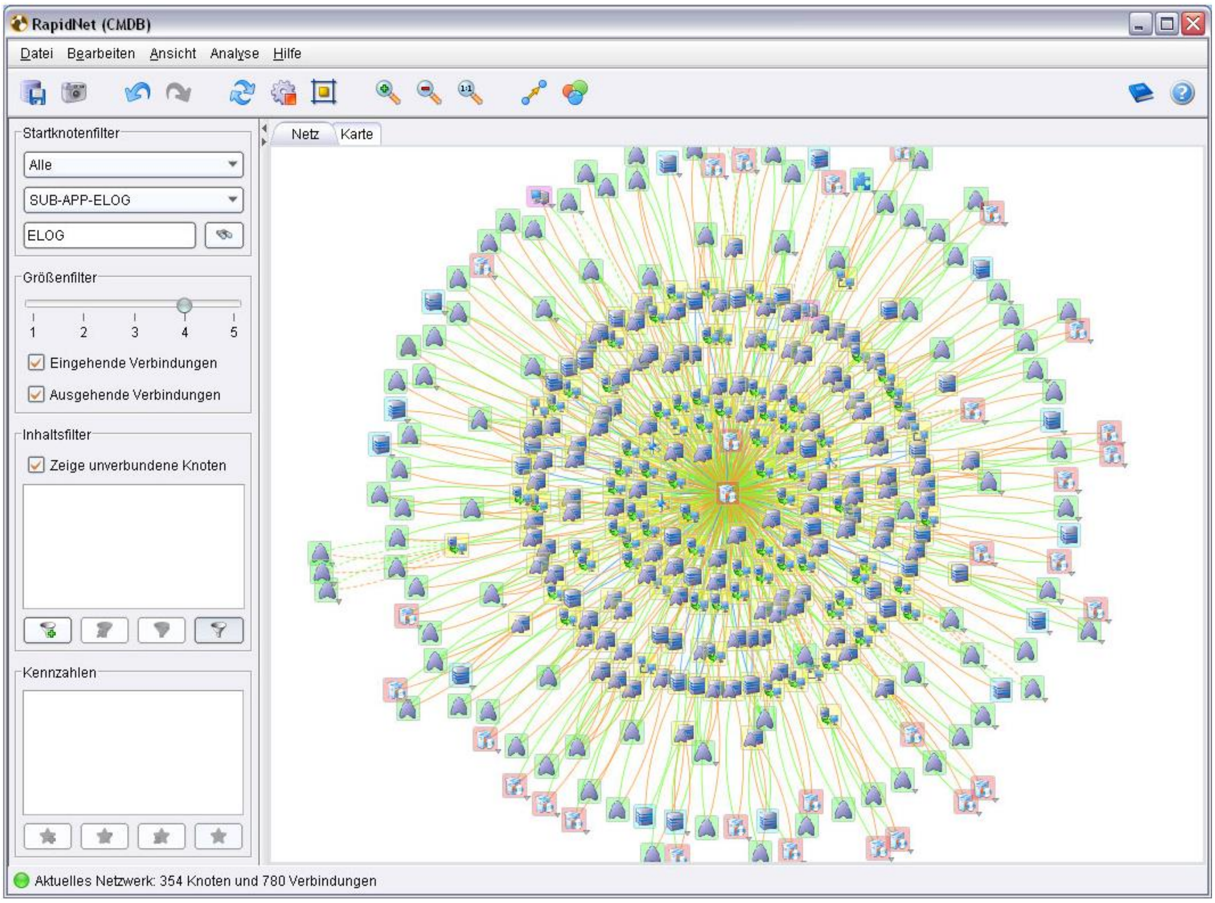
- Text Mining Extension 5.3.0
- Web Mining Extension 5.3.0

6.3.1. Structure Mining Epinions and The Product Forum

Rapid Miner does not offer any network graph displaying options. The developers "Rapid-I" offer another software based on RapidMiner on their homepage named RapidNet that addresses structural graph analysis problems. While it is not capable to preprocess data or calculate any values it should be able to display both graphs needed for testing. It not only supports displaying directed or undirected graphs but also hierarchical relations and displaying geographical information on maps.

To get this extension you have to contact the sales team. Although the support team was very friendly and interested in the purpose the software should serve it was not possible to get an evaluation version of this software for free to perform the desired tests.

Introduction and example workflows are described in the Vancouver data blog [55] as well as in [56].



Startknotenfilter

Alle

SUB-APP-ELOG

ELOG

Größenfilter

1 2 3 4 5

☑ Eingehende Verbindungen

☑ Ausgehende Verbindungen

Inhaltsfilter

☑ Zeige unverbundene Knoten

Kennzahlen

Aktuelles Netzwerk: 354 Knoten und 780 Verbindungen

Figure 6.23: RapidNet as a standalone application from the creators of RapidMiner offers data structure analysis but is not freely available for download under `http://rapid-i.com/content/view/183/193/lang,en`

6.3.2. Machine Learning based Content Mining Epinions

RapidMiner was the only mining software tested to support classifying unlabelled examples. Further it is able to classify single reviews separately and shows different results with different classification algorithms. Some experiments were performed to compare Support Vector Machines (SVM) with Naive Bayes classification.

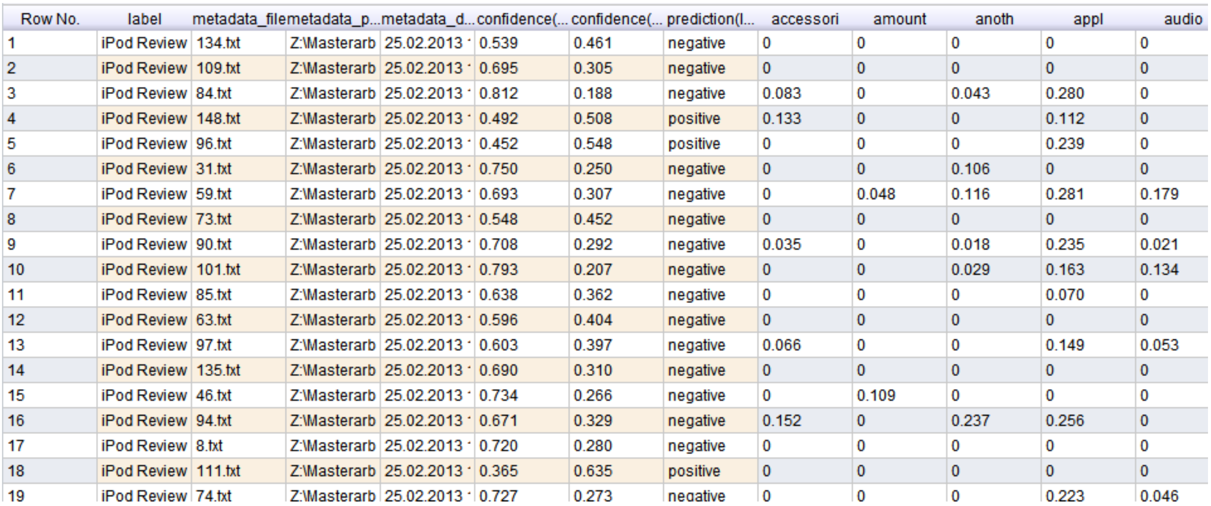
Naive Bayes as well as SVM with the SVM complexity constant c=0.0 both led to very poor results. The higher the c value the more impact an individual training sample can have on the model leading to a more specific model providing higher accuracy for a specific domain but leading to a more narrow application area. A model learned with a high c value on a PC or Laptop topic will most likely fail on other domains like movies or hotels. Setting the value to 0 only 4 of 150 reviews were classified as positive and 146 as negative. Although these 4 positives were determined correctly, in reality there were 130 reviewers recommending the iPod while only 20 gave it an overall negative rating. The overall sentiment distribution would be 2,67%/97,33% compared to the real value of 86,7%/13,3% what is way below an acceptable level.

Using SVM with c=1.0 the result increased significantly showing that a higher specialized model is better if the text documents you want to classify belong to the same domain as the training set. Still, the results were not good enough to make it suitable for accurate classification. Only 33 reviews were classified although the correct value would have been 130. The reason may be the very uneven distribution of positive and negative example sets.

Using the same settings but only 20 positives and 20 negatives as labelled examples the result improved significantly to as shown in table 6.14.

	Naive Bayes	**Support Vector Machine c=1**	**Correct value**
Overall sentiment positive/negative 40	na	48%/52%	86.7%/13,3%
Overall sentiment positive/negative 150	2,67%/97,33%	22%/78%	86.7%/13,3%
Review sentiment correctness 40	na	60,67%	100%
Review sentiment correctness 150	16%	35,33%	100%

Table 6.14: Results evaluated with both preprocessing methods. The overall sentiment was determined much too positive biased. Since classification did not work for unlabelled data, no results could be obtained when using 40 training examples.



Row No.	label	metadata_file	metadata_p...	metadata_d...	confidence(...	confidence(...	prediction(l...	accessori	amount	anoth	appl	audio
1	iPod Review	134.txt	Z:\Masterarb	25.02.2013 ·	0.539	0.461	negative	0	0	0	0	0
2	iPod Review	109.txt	Z:\Masterarb	25.02.2013 ·	0.695	0.305	negative	0	0	0	0	0
3	iPod Review	84.txt	Z:\Masterarb	25.02.2013 ·	0.812	0.188	negative	0.083	0	0.043	0.280	0
4	iPod Review	148.txt	Z:\Masterarb	25.02.2013 ·	0.492	0.508	positive	0.133	0	0	0.112	0
5	iPod Review	96.txt	Z:\Masterarb	25.02.2013 ·	0.452	0.548	positive	0	0	0	0.239	0
6	iPod Review	31.txt	Z:\Masterarb	25.02.2013 ·	0.750	0.250	negative	0	0	0.106	0	0
7	iPod Review	59.txt	Z:\Masterarb	25.02.2013 ·	0.693	0.307	negative	0	0.048	0.116	0.281	0.179
8	iPod Review	73.txt	Z:\Masterarb	25.02.2013 ·	0.548	0.452	negative	0	0	0	0	0
9	iPod Review	90.txt	Z:\Masterarb	25.02.2013 ·	0.708	0.292	negative	0.035	0	0.018	0.235	0.021
10	iPod Review	101.txt	Z:\Masterarb	25.02.2013 ·	0.793	0.207	negative	0	0	0.029	0.163	0.134
11	iPod Review	85.txt	Z:\Masterarb	25.02.2013 ·	0.638	0.362	negative	0	0	0	0.070	0
12	iPod Review	63.txt	Z:\Masterarb	25.02.2013 ·	0.596	0.404	negative	0	0	0	0	0
13	iPod Review	97.txt	Z:\Masterarb	25.02.2013 ·	0.603	0.397	negative	0.066	0	0	0.149	0.053
14	iPod Review	135.txt	Z:\Masterarb	25.02.2013 ·	0.690	0.310	negative	0	0	0	0	0
15	iPod Review	46.txt	Z:\Masterarb	25.02.2013 ·	0.734	0.266	negative	0	0.109	0	0	0
16	iPod Review	94.txt	Z:\Masterarb	25.02.2013 ·	0.671	0.329	negative	0.152	0	0.237	0.256	0
17	iPod Review	8.txt	Z:\Masterarb	25.02.2013 ·	0.720	0.280	negative	0	0	0	0	0
18	iPod Review	111.txt	Z:\Masterarb	25.02.2013 ·	0.365	0.635	positive	0	0	0	0	0
19	iPod Review	74.txt	Z:\Masterarb	25.02.2013 ·	0.727	0.273	negative	0	0	0	0.223	0.046

Figure 6.24: Part of the sentiment prediction model result for Epinions – by comparing the prediction-column with the "recommended" column in the original crawled review data table the sentiment correctness value can be determined

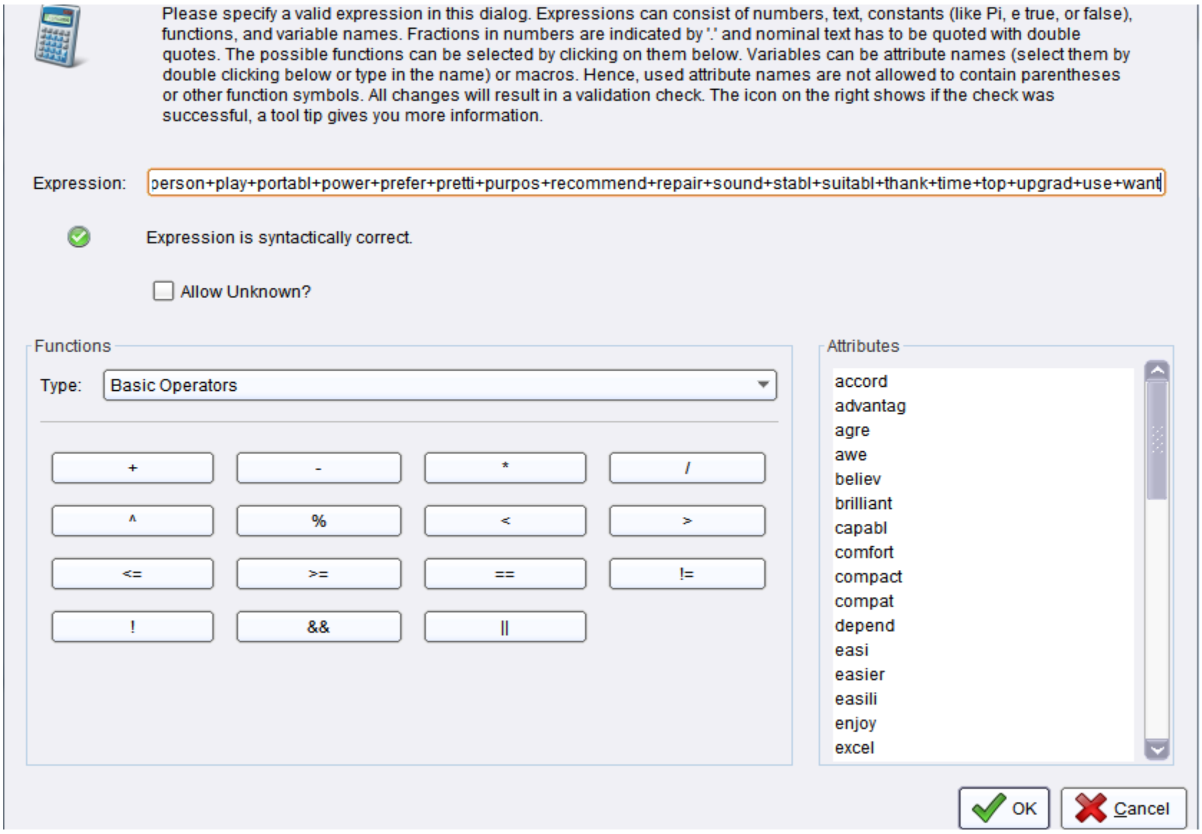
6.3.3. Lexicon based Content Mining Epinions

Extracting sentiment took more effort compared to KNIME since Rapid Miner does not support counting the number of given terms in a document directly in one node. Instead the following steps were performed to achieve the same goal:

1) Extract the transformed sentiment dictionary wordlist from KNIME and store it in a textfile. The wordlist was taken from KNIME since Rapid Miner is incapable of advanced database cell or column transformations and accepts lexica only in a 1 term per line. This textfile must be saved in MS Dos format because other formats like Unicode do not work correctly with term comparisons.
2) Import the Web Content Extractor result into the RapidMiner workflow
3) Use the Stopwords(Dictionary) node to erase any sentiment word in dictionary 1) from the review text base 2).
4) Store the resulting terms it in a text file by copying it from the result view in a text editor.

5) Now you can use the Stopwords(Directory) node again to filter all terms from the non-sentiment word list created in 3) and the result you get is the desired sentiment term list.

The content mining approach itself turned out to be more complicated than KNIME since the attribute calculations work in another way. KNIME offers the possibility to freely program small JAVA snippets that can do virtually any task with a few code lines that only needs basic programming skills. RapidMiner uses a graphical user interface to calculate values that is clearly easier to use for inexperienced users but it lacks the flexibility that java code based calculation offers. Sentiment values for a sentence were calculated by adding the corresponding line of TF-IDF values. There were a bit more than 500 terms for positive sentiment values alone. With rapid miner you have select every single terms by hand and create a formula. This would have been no problem for smaller data sets but for this task it turned out to be too much effort.



Please specify a valid expression in this dialog. Expressions can consist of numbers, text, constants (like Pi, e true, or false), functions, and variable names. Fractions in numbers are indicated by '.' and nominal text has to be quoted with double quotes. The possible functions can be selected by clicking on them below. Variables can be attribute names (select them by double clicking below or type in the name) or macros. Hence, used attribute names are not allowed to contain parentheses or other function symbols. All changes will result in a validation check. The icon on the right shows if the check was successful, a tool tip gives you more information.

Expression: person+play+portabl+power+prefer+pretti+purpos+recommend+repair+sound+stabl+suitabl+thank+time+top+upgrad+use+want

Expression is syntactically correct.

☐ Allow Unknown?

Functions

Type: Basic Operators

Attributes

accord
advantag
agre
awe
believ
brilliant
capabl
comfort
compact
compat
depend
easi
easier
easili
enjoy
excel

Figure 6.25: Interface for calculation of numeric values using RapidMiners "Generate Attribute" node with its built-in operators

To overcome the limitations regarding the calculation abilities of RapidMiner Excel was used again for TF-IDF calculations and feature sentiment determination. Using this method further improves scalability for even larger data sets.

In the end 15 features could be recognized. Surprisingly other and less features were recognized as with KNIME that found 36 features despite both programs used the same feature term dictionary. It may be the case that the indirect feature term filtering method did not work out perfectly due to the necessary format conversion and manual copying processes. This may be the reason for the smaller feature result set.

Table 6.15 shows the four features with the most positive and negative TF-IDF sentiment scores. Those features most positive were identical to KNIME in 3 of 4 cases and proved to be entirely correct. The negative features turned out to be completely different to KNIME. 2 out of 4 features proved to be actually negative.

Features found	TF-IDF value by RapidMiner	Real sentiment
sound (positive)	137,26	positive (27 positive, 21 neutral, 2 negative mentions in 1585 sentences)
screen (positive)	106,21	positive (18 positive, 27 neutral, 5 negative mentions in 915 sentences)
audio (positive)	40,63	positive (7 positive, 38 neutral, 5 negative mentions in 2843 sentences)
wheel (positive)	37,41	positive (14 positive, 31 neutral, 5 negative mentions in 2473 sentences)
output (negative)	4,37	positive (2 positive, 10 neutral, 1 negative mentions in 4452 sentences)
LCD (negative)	7,02	positive (5 positive, 16 neutral, 1 negative mentions in 4452 sentences)
switch (negative)	9,74	negative (0 positive, 17 neutral, 2 negative mentions in 4452 sentences)
power (negative)	13,40	negative (2 positive, 24 neutral, 10 negative mentions in 4452 sentences)

Table 6.15: Feature finding result showing the 4 most positive and the 4 most negative rated features compared to their manual evaluated real features.

Compared to the results in KNIME only one positive feature differs – wheel appears in the list replacing iTunes. All features were correctly classified as positive since every feature clearly had more positive than negative occurrences.

All negative features are completely different from the results obtained with KNIME. The reason may as well be the fact that none of the features recognized by KNIME (charge, customer service, warranty, batteries) appears in the list of features recognized by RapidMiner. The sentiment correctness is a bit lower being only correct in 2 out of 4 cases. "Output" has 2 positive mentions for the ability to output video on a TV and just 1 negative mention complaining about the mediocre quality of this output. LCD has an even stronger positive tendency. 5 positive mentions oppose 1 negative that complains about his display got cracked. Switch and power were classified correctly. Reviewers complained about missing on-off switch, lack of power adapter and high power drain on video playback for those features.

The same phenomena could be observed as with KNIME when looking at how often a positive or a negative term occurs in the review texts. Negative terms occurred significantly less often which means term that occur more often appear to get higher TF-IDF sentiment scores.

	Absolute value	Percentage
Precision	118/127	92,90%
Recall	119/126	94,44%
Sentiment correctness	23 correct, 21 neutral, 6 wrong	79,31%

Table 6.16: Evaluation of three metrics precision, recall and sentiment correctness using a test set of 50 sentences

Precision and recall results as shown in the table above are very close to the values that could be obtained with KNIME and are good enough to make feature recognition trustworthy. Feature sentiment correctness is almost identical to KNIME with only minor variations. Sentiment correctness is good but there is a significant amount of sentences that are classified as either positive or negative despite they express a neutral sentiment or are off topic. This could lead to a wrong biased sentiment. Overall the determined values lead to the same conclusions as with KNIME.

6.3.4. Machine Learning based Content Mining The Product Forum

Using text classification approach the results shown in table 6.17 were obtained. Contrary to Epinions the Naive Bayes classifier could deliver better results in this classification task. Compared to KNIME the overall distribution is much nearer to the actual value. The classification correctness of each post was also better than KNIME. RapidMiner could classify 20 of 26 posts correctly opposed to 19 correct posts with KNIME.

	Naive Bayes	**Support Vector Machine c=1**	**Correct value**
Overall distribution laptop/desktop 26	59,2%/40,8%	76,2%/23,8%	65,4%/34,6%
Classification correctness 26	76,9%	69,2%	100%
Classification correctness 18	72,2%	66,7%	100%

Table 6.17: Classification result using 26 or 18 posts as training and classification sets.

RapidMiner was the only program tested that allows classification of unlabelled examples. The results are accurate enough to rely on them at least if one class clearly differs from the other.

6.3.5. Lexicon based Content Mining The Product Forum

Sentiment terms had to be extracted indirectly in the same 4 steps that were used for Epinions and are described in chapter 6.3.3.

The lexicon based classification worked very well for both calculation methods. Again the better result could be obtained by using algorithm 3 which means awarding each device that has the higher value in a post exactly 1 point instead of the actual TF-IDF value. The result is almost identical to the machine learning based classification and clearly outperforms KNIME. Laptop and desktop values are 5,9% closer to the correct value when using algorithm 3.

	Laptop	**Desktop**
Overall preference distribution (algorithm 2 – TF-IDF sums)	56,2%	43,9%
Overall preference distribution (algorithm 3 – absolute preference count)	59,5%	40,5%
Correct distribution value	65,4%	34,6%

Table 6.18: Evaluation of post classification using the two calculation methods presented in chapter 6.1.6

Looking at table 6.19 quite a number of features could be found that give evidence about the pros and cons of the different device types. With the right interpretation in mind all words in the list make sense. The listed terms give a decent overview over the features that were discussed in the thread.

Manual review		Automatic review			
Laptop	**Desktop**	**Laptop pos.**	**Laptop neg.**	**Desktop pos.**	**Desktop neg.**
portability	gaming	comfortable (1,85/0,54)	break (1,00)	compatible (1,23)	fragile (1,49/1,00)
space saving	easy to exchange parts or upgrade	light (0,69)	cheap (1,00)	flexible (1,00)	expensive (3,22/1,19)
can be as powerful as a desktop (except for gaming)	cheaper repair cost	portable (2,26)		compact (0,52)	
	cheaper	stable (1,69/0,78)			
	better for designing tasks	upgrade (3,24/2,16)			

Table 6.19: Manual vs. automated feature recognition: dark green fields are determined correctly, orange fields are correct features but classified false. The values in brackets show TF-IDF(this device)/TF-IDF(other device).

4 out of 8 manually determined and actually mentioned features were recognized. These were "portable", "upgrade", "cheap" and "compact". The features found are exactly the same as with KNIME. The classification on the other hand differs quite a bit and unfortunately most features are matched to the false device or have the wrong sentiment. Only 1 out of 5 features was determined correctly (the advantage of a laptop is that it is "portable"). KNIME did a better job classifying 3 of 6 features correctly.

	Absolute number	Percentage
Number of features found	4	50%
Feature match correctness (correct device AND sentiment)	1	20%

Table 6.20: Evaluation metrics for features found

6.4. Orange

Installation: Download the “Orange with Python” file from `http://orange.biolab.si/download/`, start the installer and you are done.

Since there are some functions and extensions that are not implemented in the GUI, it can be necessary to run installation of some extensions in shell mode. Under Windows this can be done by installing the Anaconda Python distribution that automatically sets all necessary path variables and thus enables Python shell commands. You just have to enter `python <setup file name> install` in the command shell and installation takes place automatically. To install extensions further GNU Compiler Collection is required that comes with Anaconda. Nevertheless, in some cases installation of extensions refused to work. For instance, there exists a rather unpopular extension that is not integrated in the GUI named “Orange-Network”. This could have improved structure mining results but installation exited with gcc.exe error. Maybe there was a version compatibility issue. Unfortunately documentation of that addon was still in Alpha state at the time of testing and appeared to be very minimalistic. The user base giving support in the forum was very narrow. Hence errors could not be tracked back and corrected.

Function Principle: A basic difference to KNIME and RapidMiner is that nodes are executed in real-time as soon as you add them to the workflow. If you change some parameters of one node that affect functionality of other nodes you will recognize this instantly since the warning message will show up as soon as the change has been performed.

Orange offers a Python scripting interface that enables the user to program new algorithms and use the same existing data analysing procedures on a command line interface that are available from the graphical user interface.

Orange too offers a number of extensions to enhance basic functionality just like KNIME and RapidMiner. Partially they are installable from the GUI but some others can only be accessed through the Python console. The extension that could be installed and was used for testing is “text mining”. As this extension was just under construction on April 2013, it appeared to be very hard to use it. It lacks any documentation for existing nodes which means you have to use the nodes in a trial-and-error manner.

The user interface is very simple compared to KNIME and RapidMiner. It uses only one single toolbar for the node repository. Nodes are called widgets in Orange but they provide the same functionality. Node connections and connection ports are not visually distinguished as done in KNIME or RapidMiner. You just get a negative feedback, when you try to connect two incompatible ports with each other.

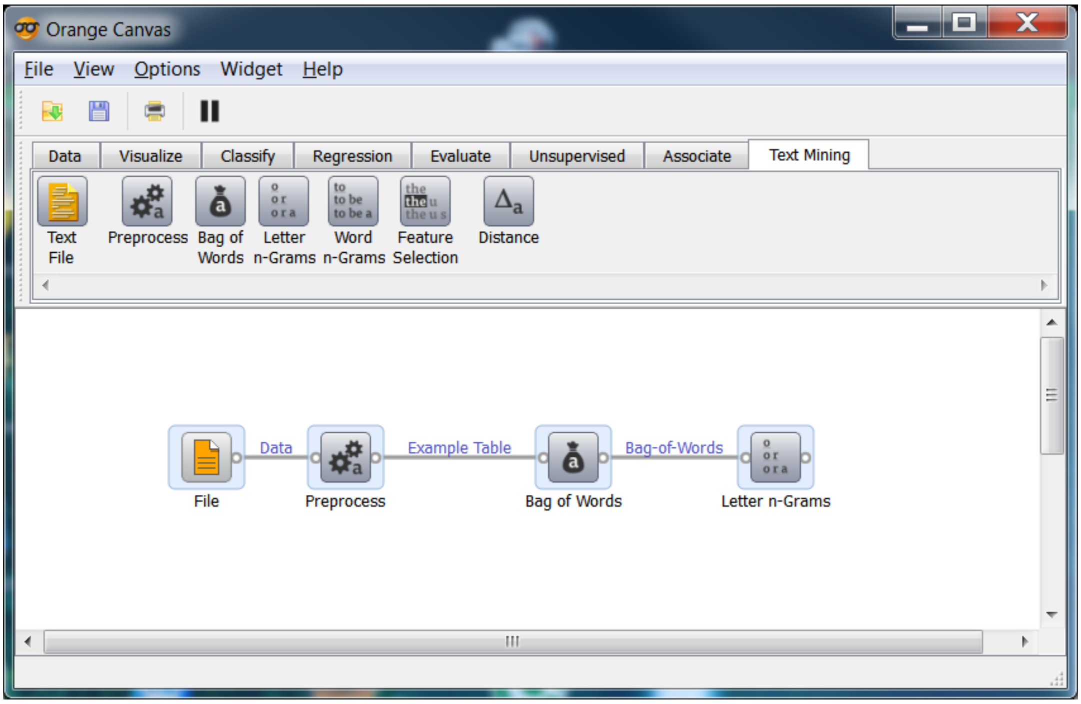


Figure 6.26: The GUI of Orange Canvas is very simple and uses up only a small amount of screen size making it perfect for smaller displays.

Besides offering a graphical user interface any command can be run by executing the underlying Python script from the command shell. A short command sequence opening a tabulator separated table and showing all feature types contained in it could look like the algorithm below.

```
import Orange
data = testfile.tab
print data.domain.features
OUTPUT: <Orange.feature.Discrete 'name', Orange.feature.Discrete 'description'>
```

Algorithm 6.4: Sample algorithm opening a tabulator separated table and displaying the cells as features “name” and “description”

6.4.1. Structure Mining Epinions and The Product Forum

Orange offers general network displaying, clustering, exploring and extraction of subgraphs. These functionalities would theoretically make the program a good choice for making connections between reviewers and post writers perfectly visible in various ways. Nevertheless test runs failed when trying to import the data table gained from the Web Content Extractor crawling process.

There are generally two ways to import data into Orange: the normal file reader and the net file reader. While the normal file reader imports the crawled table flawlessly, no table formatting could be found that let the network widgets accept the table as input. Unfortunately the Net Explorer node lacks documentation about the desired table input format.

Using the net file reader you can import a Phyton NetworkX graph but that one would have to be generated using Phython code. It should be possible to import data tables with Phython and store them in NetworkX and store it in the desired format but this requires a rather high programming effort – especially if you are not used to Phyton programming language.

Another easy to use option would have been using the pajek import format also supported by the net file reader. There exist a free software named “excel2pajek” that is able to convert .xls Excel tables into .net pajek format. Unfortunately Orange’s net reader could not read the .net files produced with this software. Since even test graphs taken from the pajek homepage refused to work with Orange one can assume that the net file reader has a bug making imports of .net files impossible.

Overall structure mining capabilities may be good in theory but are very complicated to use and buggy. The lack of documentation and small user community makes debugging extremely complicated. That is why this function could not be tested far enough to show up usable results.

6.4.2. Machine Learning based Content Mining Epinions

The general process model for classifying data with Orange is the same as for The Product Forum, and is shown in the figure below.

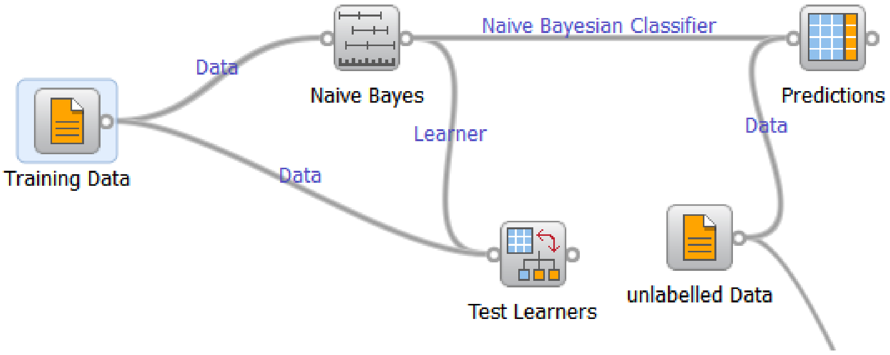


Figure 6.27: Modelled classification process using Naive Bayes classifier

Input data has to be in a tab-delimited file format with one text column followed by one classification column. The first line is the column header. The file reader node works fully automatically and allows just specifying the handling of missing values.

Unfortunately the file reader automatically recognized the text and class attributes as meta attributes when reading a data table with more greater than 20 sample lines. This made the whole classification process non-functional since the classifier could not handle that attribute format. The workaround to overcome this problem was each to use only 10 positive and 10 negative reviews as labelled examples.

When trying to import the unclassified review text list the program crashed regardless how large the data set was. The reason could not be figured out but using the whole labelled list it worked fine.

	Naive Bayes	Support Vector Machine c=1	Correct value
Overall sentiment positive/negative 40/150	na/na	na/na	86.7%/13,3%
Review sentiment correctness 40/150	na/na	na/na	100%
Overall sentiment positive/negative 20	100%/0%	100%/0%	86,7%/13,3%
Review sentiment correctness 20	86,7%	86,7%	100%

Table 6.21: Due to data import problems none of the defined tasks could be performed. Instead only 20 labelled examples – 10 positive and 10 negative – were used.

Looking at the results the overall sentiment could be determined correctly as positive but the real share was not recognized correctly. Instead all reviews were classified as positive using Naive Bayes classifier as well as SVM.

Classification did only work for one attribute and one variable that depends on this attribute. This means taking term frequency matrices or lists of tokenized terms would not lead to any better result.

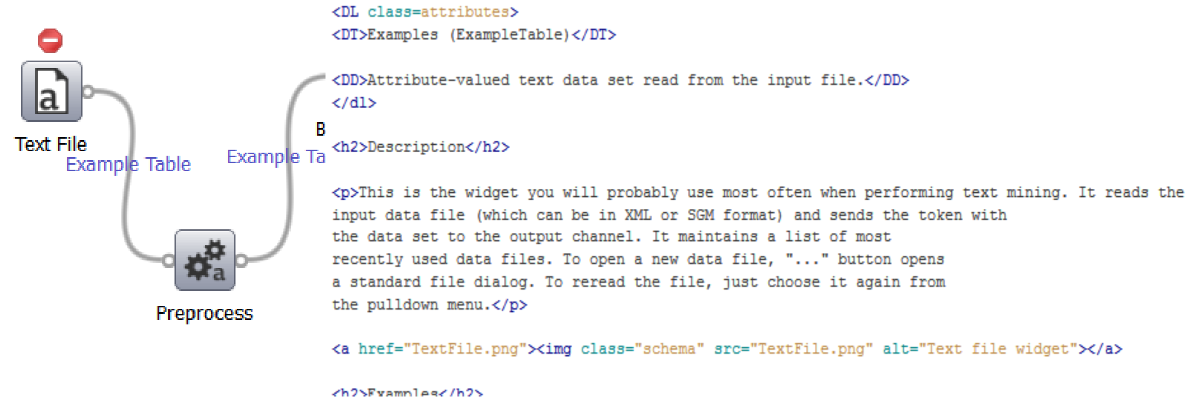
6.4.3. Lexicon based Content Mining Epinions

Orange offers a text mining extension that should allow reading text files, preprocessing, splitting documents in letter or word n-grams (tokenize them), calculate term frequency or similarity of two documents. Lexicon based term comparison is not supported at the moment.

Text mining extension was still in beta development state at the time of testing lacking any type of documentation except tooltips in the GUI and providing only limited functionality. Orange developers themselves stated in their forum that the extension had still beta status, room for improvement and they were still searching for students with Python knowledge that want to contribute to the program in March 2012[37].

During testing it was not even possible to import a text file. Since no further documentation was available, you were left on your own when trying any widget available with this extension. It was not even possible to import a text file with the “Text File” reader node. None of the formats tested - .csv, .html, .xml, .sgm, .doc, .rtf, .tab and .txt with varying charset types – were recognized. A gasp on the source code on `https://bitbucket.org` revealed that the desired input format should be XML or SGML[38] However these formats did not work on the test system. Using the standard file reader it was not possible read data and provide it to further text processing nodes.

Since all functions provided with this extension rely on the text file reader, no results could be obtained.



```
<DL class=attributes>
<DT>Examples (ExampleTable)</DT>

<DD>Attribute-valued text data set read from the input file.</DD>
</dl>

<h2>Description</h2>

<p>This is the widget you will probably use most often when performing text mining. It reads the
input data file (which can be in XML or SGM format) and sends the token with
the data set to the output channel. It maintains a list of most
recently used data files. To open a new data file, "..." button opens
a standard file dialog. To reread the file, just choose it again from
the pulldown menu.</p>

<a href="TextFile.png"><img class="schema" src="TextFile.png" alt="Text file widget"></a>

<h2>Examples</h2>
```

Figure 6.28: Although the source code of the "Text File" widget clearly states the expected .xml or .sgm format (right) the file reader refused to read any such file on the test system (left)

[37] See forum post `http://orange.biolab.si/forum/viewtopic.php?f=14&t=1516`

[38] Standard Generalized Markup Language: A markup language for documents – XML as well as HTML are based on this language ans XML has taken the place of SGML

6.4.4. Machine Learning based Content Mining The Product Forum

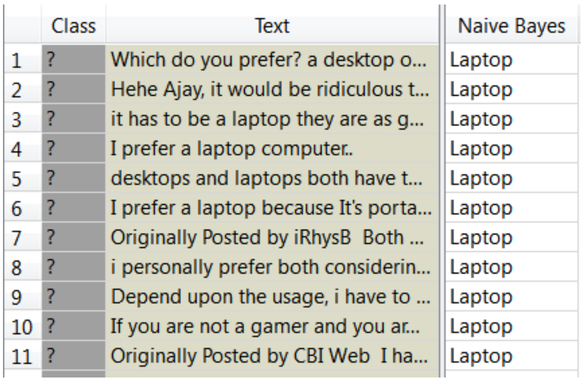
Since the same training set size limitation as with Epinions occurred tests with 26 posts was impossible and instead a reduced set of 9 desktop and 11 laptop supporting examples was used.

Classification turned out to be rather one-sided this way always favouring Laptop through the whole document set regardless if Naïve Bayes or a linear SVM classifier was used and which settings were configured. The result stayed this way even if the test learner node stated flawless classification accuracy for the model.

This resulted in a distribution of 100% Laptop and 0% Desktop which is correct when only considering the winner of the comparison but the percentage is extremely inaccurate. Hence classification correctness was more or less random.

	Naive Bayes	Support Vector Machine c=1	Correct value
Overall distribution laptop/desktop 20	100%/0%	100%/0%	55%/45%
Classification correctness 20	55%	55%	100%
Classification correctness 18	50%	50%	100%

Table 6.22: Classification result using 20 or 18 posts as training and classification sets. Note that only 20 instead of 26 posts could be classified due to data size limits of the software.



	Class	Text	Naive Bayes
1	?	Which do you prefer? a desktop o...	Laptop
2	?	Hehe Ajay, it would be ridiculous t...	Laptop
3	?	it has to be a laptop they are as g...	Laptop
4	?	I prefer a laptop computer..	Laptop
5	?	desktops and laptops both have t...	Laptop
6	?	I prefer a laptop because It's porta...	Laptop
7	?	Originally Posted by iRhysB Both ...	Laptop
8	?	i personally prefer both considerin...	Laptop
9	?	Depend upon the usage, i have to ...	Laptop
10	?	If you are not a gamer and you ar...	Laptop
11	?	Originally Posted by CBI Web I ha...	Laptop

Figure 6.29: Part of the resulting classified post list for The Product Forum

6.4.5. Lexicon based Content Mining The Product Forum

As already asserted during the test on the Epinions database Orange could neither read in the data for further text processing nor the data table itself. The text processing widgets could not handle the data read in by the standard file reader. No results could be obtained with this task.

6.5. Evaluation and Tool Comparison

The results obtained during tool tests presented in this chapter are summarized and compared in the overview tables below. Each tool got a score on the same scale from 0 to 5 points that was already used for structure mining and is described after the comparison tables. The score depends on the criteria given in the tables below. The scale is described in table 6.26.

Tool	**RapidMiner**	**KNIME**	**Orange**
Homepage	`http://rapid-i.com/content/view/181/190/`	`http://www.knime.org/`	`http://orange.biolab.si/`
Structure Mining Epinions	na	☺☺☺☺☺	☹
Summary	no test could be performed; needs RapidNet extension that is not available for free	Complete and well displayed result graph	Graph import function did not work using the graphical user interface. No format tested could be read.
Graph can be displayed	No	Yes	No
Graph can be filtered	No	Yes	No
Graph error count	na	0	na
Structure Mining The Product Forum	na	☺☺☺☺☺	☹
Summary	no test could be performed; needs RapidNet extension that is not available for free	Complete and well displayed result graph	Graph import function did not work using the graphical user interface. No format tested could be read.
Graph can be displayed	No	Yes	No
Graph can be filtered	No	Yes	No
Graph error count	na	0	na
Content Mining Epinions – Machine Learning based	☺☺	☹	☺
Summary	The only classifier that could classify each single review. Unfortunately the overall sentiment was not determined correctly.	No classification of unlabelled data can be done. The overall sentiment could be determined correctly only with fully labelled examples.	Overall sentiment was classified correctly but no distinction of single reviews was possible. More than 20 training sets could not be used.
Classification of unlabelled data possible	Yes	No	Yes
Classification of each single review worked	Yes	Very Limited	No
Overall sentiment positive/negative 40 (deviance to correct value)	48%52% (38,7%)	na	100%/0% (13,3%) (only 20 labelled examples)
Overall sentiment positive/negative 150 (deviance to correct value)	22%/78% (64,7%)	99,3%/0,7% (12,6%)	na
Review sentiment correctness 40	60,67%	na	86,67% (only 20 labelled examples)
Review sentiment correctness 150	35,33%	86,7%	na

Table 6.23: Comparing mining tool functionality and results - part 1.

Tool	RapidMiner	KNIME	Orange
Content Mining Epinions - Lexicon based	☺☺	☺☺☺☺	☹
Summary	worked only with complex indirect term filtering method and with support of Excel; less features were recognized than in KNIME	KNIME provides the most complete solution. The whole process could be satisfactory performed within the program.	Orange text mining addon was still in beta state, lacked any documentation when this thesis was written and refused to read any text file.
Process can be done with the program alone	No	Yes	No
Term filtering worked directly and convenient	No	Yes	No
Positive features found	**sound** (correct) **screen** (correct) **audio** (correct) **wheel** (correct)	**sound** (correct) **screen** (correct) **iTunes** (correct) **audio** (correct)	na
Negative features found	**output** (wrong) **LCD** (wrong) **switch** (correct) **power** (correct)	**charge** (correct) **customer service** (wrong) **warranty** (correct) **batteries** (correct)	na
Precision	92,90%	85,95%	na
Recall	94,44%	96,36%	na
Sentiment correctness	79,31%	81,48%	na
Content Mining The Product Forum – Machine Learning based	☺☺☺☺	☺	☺☺
Summary	Classification could be done in a very convenient and accurate way with the right classifier (Naive Bayes did the best job).	Classification could only be done on labelled posts. Acceptable results were only achieved when using even distributed training examples.	Overall classification worked out for unlabelled examples but accuracy was very poor. More than 20 training sets could not be used.
Classification of unlabelled data possible	Yes	No	Yes
Classification of each single review worked	Yes	Limited	No
Overall distribution laptop/desktop 26 (deviance to correct value)	59,2%/40,8% (6,2%)	92,3%/7,7% (26,9%)	100%/0% (34,6%) (only 20 labelled examples)
Classification correctness 26	76,9%	73,1%	55% (only 20 labelled examples)
Classification correctness 18	72,2%	66,7%	50%

Table 6.24: Comparing mining tool functionality and results - part 2.

Tool	**RapidMiner**	**KNIME**	**Orange**
Content Mining The Product Forum - Dictionary based	☺☺☺	☺☺☺☺	☹
summary	Mining delivered acceptable yet imprecise results. For table transformations an external spreadsheet program like Excel was obligatory.	General classification was done accurately and feature classification accuracy was best in the test field.	The same problem as with dictionary based content mining on Epinions occurred: the program refused to read any text file.
process can be done with the program alone	No	Yes	No
term filtering worked directly and convenient	No	Yes	No
dictionary based classification (deviance to correct value)	59,5%/40,5% (5,9%)	53,6%/46,4% (11,8%)	na
positive features found for laptop	**portable** (correct) **upgrade** (wrong)	**portable** (correct) **compact** (correct) **upgrade** (wrong)	na
negative features found for laptop	**cheap** (wrong)	**expensive** (correct)	na
positive features found for desktop	**compact** (wrong)	**compact** (wrong)	na
negative features found for desktop	**expensive** (correct)	**cheap** (wrong)	na
percentage of features found	50%	50%	na
feature match correctness	20%	50%	na

Table 6.25: Comparing mining tool functionality and results - part 3.

☹	No (usable) results could be achieved
☺	Some results could be achieved but they are very faulty and incomplete.
☺☺	Bad result quality, some minor result parts are missing.
☺☺☺	Acceptable result quality with moderate errors, all tasks could be completed, maybe with support from other tools
☺☺☺☺	Very good and complete results with only minor errors that could be achieved with the program alone
☺☺☺☺☺	Flawless result with (nearly) no errors that could be achieved with the program alone

Table 6.26: Scale used to grade mining tools.

6.6. Choosing best Tools

Unlike the crawling tools for mining tasks there is no single program that could provide the best results for all mining tasks. Instead RapidMiner and KNIME could each show their strengths in different areas.

KNIME proved to be the only choice for structure mining on both communities. It is able to display the graph correct and to its full extent and allows filtering the graph in a way that focuses the view on the essential nodes and edges. Further it is possible to transform and prepare data tables entirely within the program.

KNIME managed to stand out from the other programs when performing lexicon based content mining tasks. Classification could be done a tad better than with the second best candidate RapidMiner. Again KNIME had the advantage that it was the only program to do the whole mining process including any data table transformations within the program. Additionally it could perform the process with less nodes than RapidMiner mainly because term filtering worked much easier and could be done in one single node.

RapidMiner could show its strength when it came to machine learning based content mining. It was the only program that could handle the task as desired. KNIME could not classify unlabelled data making it practically unusable. Orange could not be used with more than 20 labelled examples and could not distinguish between single reviews or posts individually allowing only a rough overall classification. RapidMiner was easy to use, could classify single reviews or posts separately and showed good classification accuracy in The Product Forum. Overall classification accuracy was worse when used for Epinions and failed to prefer the correct device. Single review classification worked in any case.

With a combination of KNIME for structure mining and lexicon based content mining tasks and RapidMiner for machine learning based content mining good or at least acceptable results could be achieved for any task.

7. Discussion of Results

Using the knowledge gathered through literature research and various software tests we are now able to answer the questions defined in chapter one:

1. **What type of information interesting for market research purposes could you explore with web mining tools?**
   In the literature reviewed basically three different kinds of information are given which can be obtained through web crawling and mining:
   - Structure Mining shows the relationships between users that communicate in a forum or rating platform where they rate each other. This type of mining result is shown in a structure graph where the vertices represent users and the edges connecting them represent either communication or rating lines.
   - Text Classification Content Mining allows assigning text documents to two or more different classes automatically. These text documents can be reviews on a product rating platform or forum texts discussing pros and cons of a certain product. Text classification is working with supervised machine learning methods like a Naive Bayes learner using a set of pre-classified example text documents to automatically classify a larger set of similar documents. This process results in labelling any text document either as positive or negative in reviews on a rating platform or as a vote in favour or against a certain product in a product forum discussion.
   - Lexicon Based Content Mining can be used for finding key terms in text documents such as product features and corresponding sentiment terms. This approach requires either pre-defined dictionaries or dictionaries automatically created using an algorithm on some seed words or training text sets. By counting co-occurrences of feature and sentiment terms the different features can be classified as pros or cons forming a list of positive and negative characteristics of a product. Lexicon based content mining can also be used to classify whole text documents similar to the text classification approach.
2. **Which communities are suitable for web mining data sources?**
   Online communities should meet the following criteria to qualify for providing information for a specific topic or question:
   - The community is focused on the relevant topic
   - It has as much traffic as possible
   - It houses a multitude of discrete message posters
   - Topic-related posted data is rather detailed
   - Interaction between members is referring to the research question

   To make a community suitable as data source for automated web crawling and mining purposes the following characteristics should be met:
   - Information is consistent in making uploaded historical data visible
   - Information is freely available without registration
   - The preferred site language is English due to dictionary word recognition problems with other languages.
   - For structure mining purposes either a user rating system is applied or direct interaction between users takes place
   - A site is suitable for content mining if users either directly rate and review products or discuss a specific product.

Excluding all community types that do not meet those criteria leaves following types of communities suitable for web mining:

- Product rating platforms
- Topic oriented communities
- Public blogs
- Some but not all commercial communities

Whilst the first three community types are clearly distinguishable from each other, commercial communities can contain virtually any type and just add the money generating factor to it. This thesis selected the product rating platform "Epinions" and the topic oriented community "The Product Forum" as base for further practical crawling and mining tests. Public blogs and the corresponding comments can contain valuable information about product market reception as well but were left out since data extraction and evaluation had worked basically the same way as with a forum where the blog text corresponds to the opening post and the comments to the answering posts.

3. **Which information can you gather through web crawling and mining?**
   To ensure a comprehensive testing of crawling and mining tools under review, for each of the two chosen communities three questions were defined – one for each data mining type.

   The questions that asked about **Epinions** are:

   - **Structure Mining:** Given a specific review about a product – how trustworthy is the reviewer? Is he given trust by independent, trustworthy members?
   - **Text Classification Based Content Mining:** Which reviews are positive and which ones are negative? What is the share of positive or negative reviews?
   - **Lexicon Based Content Mining:** For a given product – what are the positive and negative features mentioned in the product reviews?

In order to answer the structure mining task a directed graph displays a user as a vertex connected by an edge pointing to another vertex showing that the first user trusts the second one. Text classification mining leads to a list of complete reviews either labelled as positive or negative. This list shows the percentage of positive and negative reviews. The lexicon based methods output is a list of product features that are each classified either as positive or negative.

   The questions asked about **The Product Forum** are:

   - **Structure Mining:** How strong are forum members linked to each other? Who communicates with whom and to what extent?
   - **Text Classification Based Content Mining:** What is the share of laptop and desktop supporters?
   - **Lexicon Based Content Mining:** What are the advantages and disadvantages of laptops or desktops mentioned in the posts?

   The structure mining question stated above can be answered by creating a directed graph where vertices are forum members and the directed edges show who wrote a message to whom. Text classification leads to a ratio of laptop supporters to desktop supporters. The lexicon based method provides a list of product advantages or disadvantages of laptops and desktops.

4. **Which available tools are suitable for crawling and mining data necessary to answer the questions defined in the first place?**

In order to make a tool suitable for the desired crawling and mining task it has to fulfil a number of requirements. If it fulfils only one or two of these requirements it helps answering some questions. A program that fulfils all requirements can be used to perform the full web mining process without using other tools for subtasks:

- The tool should support web crawling like collecting data from websites and storing them in databases or data tables.
- It should support structure mining enabling to display data sets and their relationships as graphs.
- It should support text mining to analyse and process plain, unstructured texts and make the content machine readable.

Following requirements are set for this thesis in order to concentrate on those tools most advanced and convenient to use:

- They must run stand-alone excluding programming frameworks in order to avoid extensive programming and focus on tool appliance
- They must run on Windows because these programs offered more functionality and superior user friendliness during tests in chapter 4.3.

An extensive but not necessarily all-embracing analysis of software crawling and mining tools revealed a selection of freeware tools that met the criteria above:

Free to use crawlers:

- Newprosoft Web Content Extractor (trial version limited to 14 days and 150 records)
- Newprosoft Winweb Crawler 2.0 (trial version limited to 15 days)
- RapidMiner (trial version limited in service and extension functionality and commercial use is forbidden)

Free to use miners:

- KNIME (Open Source)
- Orange (Open Source)
- RapidMiner (trial version with the same restrictions as the RapidMiner crawler)

Commercial software offers a wide range of functionality and any program analysed would meet the requirements as well. Since prices for commercial software can be as high as 10.000 to 20.000 Euros, this type of software is only profitable for companies making professional use of the data analyses results. Commercial software was excluded from the practical testing phase. Whether you need commercial software depends mainly on the data set size you want to analyse and the support you need. Commercial software has the advantage of reliable support and mostly higher data set capacities. Commercial tools meeting the requirements are Angoss Knowledge Studio, Coheris SPAD, RapidSentilyzer, SAS, SPSS Modeller and Oracle Data Mining.

5. **How do existing data mining tools prove themselves in real-world application? How do they compare to each other?**

The experiences made during practical testing in chapter 5 are summarized as follows regarding the functionality and differences between the tested crawling tools:

Winweb crawler 2.0 allows specifying a URL, the retrieval depth and URL pattern that the crawler should follow or should ignore. Extracting the source code of a site is easy but postprocessing of the gained data is necessary to extract the relevant text parts. The links to follow-up can only be defined roughly making automatic structure extraction impossible when higher complexity is required like in Epinions.

Pros: small program size

Cons: poor quality of results (cannot extract certain text fragments, too much noise), following links did not work during test

RapidMiner offers a crawling function within its mining workflow environment. The crawler offers sophisticated links to follow-up and content extraction techniques allowing to specify URL-patterns and link text patterns. The structure crawling works by string and substring matchup as well as XPath HTML content extraction. Theoretically, it allows extracting relevant text parts. Unfortunately, the practical test results could not hold up to the theoretical abilities of the program since XPath or String content extraction did not function properly and the Epinions user profile sites necessary for structure crawling could not be read.

Pros: lots of options to extract precise text parts from a webpage

Cons: requires XPath programming skills, data extraction would not work in some cases and weblinks did not work at all

Newprosoft Web Content Extractor is the easiest tool to use of the three – especially if you are not into programming. It displays the website, allows you to mark text passages you would like to extract and weblinks. It automatically creates the underlying data extraction code in its proprietary language which can be manually altered afterwards. Following the weblinks as well as the text content extraction worked both surprisingly well and led to the best result of the three crawling tools tested.

Pros: graphical user interface suitable for less advanced users, best results of all tested tools

Cons: sometimes faulty extraction and link-follow-up that could be corrected by manual code editing in most cases

The miners tested in chapter 6 had a more common function principle than the crawlers and work with very similar user interfaces. They all share the same type of workflow modelling which includes nodes and edges. Nodes represent execution tasks like reading a source file, transform the table or count the term frequency and edges represent the data flow between those tasks. The exact differences are described below:

KNIME turned out to be the most complete solution when used for lexicon based content mining. It was the only program that could retrieve well usable results for both structure and lexicon based content mining without the need of any additional calculation or spreadsheet program. It has built-in table transformation functionality as well as word tagging. That allows searching words from a lexicon in documents, counting their term frequency and displaying the results in graphical and tabular forms. Text classification on the other hand cannot be done satisfactorily since no classification of unlabelled data is supported which makes it only suitable for theoretical model building.

Pros: comprehensive tool that can do all necessary data processings and transformations for both structure and text mining, very good results

Cons: text classification not supported

RapidMiner has one major difference to KNIME that comes in mind when using it for the first time. In KNIME the whole workflow can easily be modelled in one large task sequence graph since every node can be executed separately and any intermediate result can be watched at any time. In RapidMiner the whole workflow is executed at once and only the end result can be seen. This makes splitting the workflow into several pieces necessary for larger workflows. The second major difference is the inability to conduct advanced table transformation tasks which makes the use of an external spreadsheet program necessary. The lexicon based data mining could not be conducted in the same easy way as in KNIME since most calculations had to be made in an external spreadsheet program. Lexicon word comparison could not be done directly in one single node, it needed instead 2 full workflow processes. On the other hand RapidMiner hands down beats KNIME when it comes to do text classification on a learning model base. The built-in classifier nodes are as easy to use as in KNIME, they worked more accurately during the test and allowed labelling of unlabelled examples in a sufficiently accurate manner. Displaying structure mining result graphs is supported only with the not freely available extension RapidNet making the free version incapable of doing this.

Pros: good text classification results, content mining produces acceptable results

Cons: no structure mining with the free version, external table data transformation necessary, content mining possible but more tedious and worse results than KNIME

Orange is the youngest and least mature project of the three mining tools under review. It works in two ways – either you can use the graphical user interface or you can operate the Python function nodes from a command shell. Either way there were significant problems reading in data graph models and content for the text mining module which was available in Beta state only at time of testing. This made structure mining as well as lexicon based content mining functionality practically unusable. Text classification on the other hand worked well with an easy to model workflow that offered acceptable result accuracy when used to classify unlabelled examples. What makes the program stand out from the others is its simple and space-saving user interface that manages to display everything simultaneously (e.g. node description, node repository, workflow project, console). It shows only available nodes and the workflow making it ideal for smaller screen sizes.

Pros: simple and clearly arranged user interface, text classification produces acceptable results

Cons: cannot read test data correctly, Beta state with some options missing, text classification possible but lower result quality than RapidMiner

6. **How precise are the results? Do the tools gather information on a site correctly or not?**

Of the three text mining procedures tested **lexicon based opinion mining** turned out to be the most interesting type. It not only led to the most significant results but also delivered the highest accuracy of all tests. While KNIME and RapidMiner both delivered different results regarding the most positive and the most negative features of an iPod, both recognized the sentiment of those features correctly in 6-7 of 8 cases giving a satisfying 75-87,5% hit rate.

To get good results from this method great attention had to be paid to the feature lexicon. The result is highly dependent on the feature terms chosen and what meaning you allocate to them. For instance, the term “video” turned out to be completely inappropriate to give evidence about the video reproducing quality of the device. Most times it was used only to call the device as “ipod with video” or mention video formats the device is capable to playback. Choosing the right meaning of a term in an actual context will make the difference between a correct or false result.

Using no specific feature lexicon as it was done with The Product Forum turned out to decrease feature classification accuracy to about 50% making the results close to at random. The only exception was the dictionary based classification of short posts in The Product Forum when taking into account word order. Using no lexicon can nevertheless give a decent overview about the features discussed in the analysed text by displaying frequently used terms.

The **text classification** based approach worked well in determining the overall sentiment across the whole test set. But none of the tools could show sufficient accuracy when classifying each single review in Epinions or each single post in The Product Forum correctly. For instance, in Epinions all 150 reviews were classified as positive although 20 of them should be negative making the distinction of positive and negative reviews impossible. Another tool classified 72 reviews as positive when there should be 130 positive ones, making the result very inaccurate.

Structure mining worked overall very well in general and gave perfectly accurate results when using the correct tools (Web Content Extractor and KNIME). The most challenging task was crawling the demanded links to get appropriate source data material from which the tool could derive a well interpretable graph.

7.1. How do the results compare to other publications mentioned in related work?

Structure mining results compare well with those graphs presented in [16] and [19]. The result graph from [16] shows better edge labelling when compared to KNIME’s graph and displays messages to their full extent. [19] presents a more complex result graph. Instead of connecting user nodes directly with each other they are linked with their interests that are also represented as nodes. Users sharing the same interest appear as indirectly linked over the interest node. The graph considers a time dimension as well by displaying every interest as a multicoloured bar where every colour represents user activity during a time period.

Machine learning based text mining results can be compared with [25] and [27]. The results there are close to those obtained within this thesis. Result accuracy in [25] is between 72,8% and 82,9% depending on the lexica and classification algorithm used opposed to the best value of 86,7% achieved by KNIME on Epinions. The major difference of the proposed method is that no manually labelled examples were used. Instead they were trained automatically. [27] classifies short messages in Twitter.

Classification accuracy is slightly above 80% using the Naive Bayes algorithm showing that for domain-specific classification accuracy around 80% can be considered as normal.

The range of **lexicon based content mining** papers is the widest. Classifying whole reviews is done in [20], [21], [26] and [22]. [20] classifies mixed reviews of different products by using valence shifting. Sentiment classification correctness lies between 65% and 80% depending on the valence shifting method used. RapidMiner and KNIME could achieve even better values of 79-81% without any valence shifting. The tools probably scored that high because all reviews belonged to the same domain. Note that the values cannot be directly compared since the sentiment correctness in this work is checked for every single sentence while in [20] full reviews are evaluated. [21] uses an automatically created dictionary to achieve a classification correctness of 82,6% for positive and 52,4% for negative documents. It shows the same characteristic encountered during feature classification in this work: positive features were classified much more precise than negative. [26] performs a two-part process. First polar expressions are selected and then their sentiment is determined. Classification accuracy is around 63% - clearly lower than the 79-81% shown in this thesis. Since the statements classified in [26] are not domain-specific, this is nothing to wonder about. Surprisingly the paper shows higher precision and recall values for negative statements than for positive – the exact opposite discovered during this work and shown in [21]. [22] classifies documents using valence shifters and by counting the number of positive and negative terms. Result accuracy is up to 69,3% with precision around 70% and recall around 80%.These are about on par with other methods classifying non-domain specific reviews and clearly lower than the domain-specific classification in this thesis (accuracy: 79-81%, precision: 86-93%, recall: 94-96%). The results cannot be directly compared since single sentence precision/recall cannot be compared with full document classification. Besides neutral sentences were not counted for this thesis but in [22] they were.

Extracting and classifying single features from the review text corpora was done in [23] and [24]. [23] determines product features using association mining to find commonly used feature terms in reviews. Adjectives next to feature terms are considered as opinions. Using this method on reviews for 5 different products resulted in average precision of 64,2%, recall of 69,3% and sentence orientation accuracy of 84,2%. While precision and recall seem lower than those determined within this work (both above 90% in Epinions), you have to keep in mind that neutral sentences were not counted as failures for this work leading to but in [46] they were. Sentence orientation accuracy is similar to this work and lies around 80%. [24] uses ontologies to identify key features. Once more the nearest adjectives are considered as feature opinions. The tool shows a very low rate of misclassified features around 5%. This is a remarkable good rate when compared to feature evaluation for Epinions that showed an error rate of 12,5% to 25% for the 4 best and worst features. Using only those adjectives nearest to the features term seems to improve result accuracy compared to using all adjectives at once.

8. Summary and Future Outlook

This thesis gave an overview over available web crawling and data mining tools as well as the practical abilities of freeware tools. The major goal was to find free software suitable for conducting web mining in online communities and to test each suitable program on community websites.

Through literature research similar projects were examined and targets for the mining process were defined. Two communities appropriate for web data mining were chosen: The rating community “Epinions” and the topic oriented community “The Product Forum”. On each of this community one crawling task and three mining tasks were performed - one task per data mining type. These three mining types are structure mining, text classification mining and lexicon based content mining. Structure mining displays relationships between community users, text classification mining divides text in positive or negative groups and lexicon based content mining discovers features and whether they appear in a positive or negative context.

After setting the goal three crawling and three mining tools were selected according to the specific requirements given. These tools were used to crawl data from Epinons and The Product Forum and to fulfil the three tasks per community. The results were two directed graphs for the structure mining task, two ratios of positive and negative ratings or user postings for the text classification task and two lists of positive and negative product features for the lexicon based content mining task.

Results show that web mining in online communities is possible with freeware web mining tools. Combining the crawling abilities of Newprosofts Web Content Extractor with the text classification abilities of RapidMiner and the lexicon based content mining and structure mining functions of KNIME acceptable results could be achieved for all three mining tasks. Other tools have shown to fulfil tasks to some extent as well but some tools were not tested at all because they did not meet essential criteria or could not be used for free.

Most accurate results were achieved with web structure mining showing a perfectly accurate result graph. Lexicon based text mining still provided good results with an accuracy of 7 out of 8 correctly classified features when a predefined product feature lexicon was used. Text classification could be done to determine the overall ratio of positive and negative texts but was not able to classify each single text correctly.

It turned out to be crucial for text classification mining to use a predefined feature lexicon, otherwise the accuracy of the mining result would drop to random level. In order to achieve accurate results it is required knowing the features you want to extract and evaluate as well as interpreting them correctly.

This thesis tried to give a comprehensive overview over the performance of web mining in online communities and the abilities of tools currently on the market. Nevertheless, further research domains have to be investigated and processes can be refined according to particular demands.

First the result accuracy of content mining could be enhanced by adapting the sentiment lexica used for determining the sentiment bias by fostering the focus on the domain under review (in this case mp3 player device reviews and computers as well as laptops). The feature lexicon may also be further refined to improve the precision of feature evaluation for a specific domain. A method presented in chapter 2 using a search machine to extract domain-specific terms and considering polarity changing words provides superior results compared to a predefined lexicon [21]. Result accuracy could further be improved by taking into account negotiations as “not” and intensifiers as “very” or “somewhat”. This has been proved in [20].

During tests within this work, stemming showed to improve results especially when analysing shorter text parts or classifying text parts with lexicon based methods. Introducing stemming in future web mining approaches will therefore very likely lead to better results. Besides already available stemming methods mentioned in this thesis there are further advanced stemming methods under development. These take into account syntactic and semantic context to minimize the probability of over- and understemming leading to more accurate results [50].

This thesis focussed on practical tests of tools available free to use only. As a consequence the probably most advanced tools mentioned under “commercial tools without trial option” could not be included in the practical test phase. Further research can be done by testing educational or full versions of those programs provided the required financial resources are granted.

An interesting task would be investigating in detail how the results scale when the tools are used on larger data volumes. While Web Content Extractor worked very well for data extraction from a website, it was only available as trial version with a data size limit of 150. It was impossible to crawl and mine larger data sets in order to see if larger data volumes remain manageable and if result quality changes with larger projects.

9. References

[1] Robert v. Kozinets (2002). *The Field Behind the Screen: Using Netnography for Market Research in Online Communities.* Journal of Marketing Research, 39, pp. 61-72

[2] Walter Steinbach (1997). *Data Mining-Software (Leistungsmerkmale, Hersteller, Nutzer).* Term paper, Leipzig university

[3] Zaiane, O.R. (1998). *Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs.* Research and Technology Advances in Digital Libraries 1998

[4] Alberto H.F. Laender, Berthier A. Ribeiro-Neto, Altigran S. da Silva, Juliana S. Texeira (2002) *A brief survey of web data extraction tools.* ACM SIGMOD Record Volume 31, Issue 2

[5] John F. Elder IV & Dean W. Abbott (1998). *A Comparison of Leading Data Mining Tools.* Fourth International Conference of Knowledge Discovery & Data Mining, New York

[6] Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi (2013). *Overview of Web Content Mining Tools.* The International Journal of Engineering And Science (IJES) ISBN: 2319-1805

[7] Naoko Oyama, Yoshifumi Masunaga, Kaoru Tachi (2006). *A Diachronic Analysis of Gender-Related Web Communities Using a HITS-Based Mining Tool.* Springer-Verlag, Berlin Heidelberg, from APWeb 2006, pp. 355-366, LNCS 3841

[8] Remy Cazabet, Maud Leguistin, Frederic Amblard (2012). *Automated Community Detection on Social Networks: Useful? Efficient? Asking the users.* WI&C'12, Lyon, France, ACM 978-1-4503-1189-2/12/04

[9] R. Cazabet, F. Amblard (2011). *Simulate to detect: a multi-agent system for community detection.* Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International.

[10] Mikolaj Morzy (2009). *Mining Social-Driven Data.* Institute of Computing Science, habilitation thesis

[11] Mikolaj Morzy, Juliusz Jezierski (2006). *Cluster-Based Analysis and Recommendation of Sellers in Online Auctions.* Lecture Notes in Computer Science, Volume 4083/2006, pp 172-181

[12] Mikolaj Morzy (2005). *New Algorithms for Mining the Reputation of Participants of Online Auctions.* Springer, LNCS 3828, pp. 112-121

[13] Mikolaj Morzy, Adam Wierzbicki (2006). *The Sound of Silence: Mining Implicit Feedbacks to Compute Reputation.* Springer, LNCS 4286, pp. 365-376

[14] Mikolaj Morzy, Marek Wojciechowski, Maciej Zakrzewicz (2005). *Intelligent Reputation Assessment for Participants of Web-Based Customer-to-Customer Auctions.* Springer, LNAI 3528, pp. 320-326

[15] Mikolaj Morzy, Adam Wirzbicki, Apostolos N. Papadopoulos (2009). *Mining online auction social networks for reputation and recommendation.* `http://www.cs.put.poznan.pl/mmorzy/` (Aug 2012)

[16] Anatoliy Gruzd, Caroline Haythornthwaite (2008). *The Analysis of Online Communities using Interactive Content-based Social Networks.* `https://www.ideals.illinois.edu/bitstream/handle/2142/11505/AM08_gruzd_hay_ICTA.pdf?sequence=2` (Aug 2012)

[17] Diana Maynard, Kalina Bontcheva, Dominic rout (2012). *Challenges in developing opinion mining tools for social media.* `http://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf` (Aug 2012)

[18] Meeyoung Cha, Alan Mislove, Krishna P. Gummadi (2009). *A Measurement-driven Analysis of Information Propagation in the Flickr Social Network.* ACM, ISBN: 978-1-60558-487-4, pp. 721-730

[19] Dieudonné Tchuente, Marie-Francoise Canut, Nadine Baptiste Jessel, André Péninou, Anass El Haddadi (2010). *Visualizing the evolution of users' profiles from online social networks.* IEEE Xplore Digital Library, E-ISBN: 978-0-7695-4138-9, pp 370-374

[20] Maite Taboada, Julian Brooke, Milian Tofiloski, Kimberly Voll, Manfred Stede (2011) *Lexicon-Based Methods for Sentiment Analysis.* Computational Linguistics Vol. 37, No. 2 pages 267-307

[21] Ali Harb et al. (2008). *Web Opinion Mining: How to extract opinions from bolgs?* Published in CSTST'08: Inernational Conference on Soft Computing as Taransdisciplinary Science and Technology.

[22] Kennedy, Alistair and Diana Inkpen (2006). *Sentiment classification of movie and product reviews using contextual valence shifters.* Computational Intelligence, 22(2):110-125

[23] Minqing Hu, Bing Liu (2004). *Mining and Summarizing Customer Reviews.* ACM 1-58113-888-1/04/0008

[24] Dwi A.P. Rahayu, Shonali Krishnaswamy, Cyril Labbe, Oshadi Alhakoon (2010). *Web Services for Analysing and Summarising Online Opinions and Reviews.* Springer, LNCS 6481, pp. 136-149

[25] B. Pang, L. Lee, and S. Vaithyanathan. (2002). *Thumbs up? Sentiment classification using machine learning techniques.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86.

[26] T. Wilson, J. Wiebe, and P. Hormann (2005). *Recognizing contextual polarity in phrase-level sentiment analysis.* Proceedings of Conference on Empirical Methods in Natural Language Processing, Vancouver

[27] Alec Go, Richa Bhanyani, Lei Huang (2009). *Twitter Sentiment Classification using Distant Supervision.* Stanford Computer Science, `http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf`

[28] Liu, Bing (2007). *Web data mining: exploring hyperlinks, contents and usage data.* Springer, 2007 ISBN 978-3-540-37881-5

[29] Jesse Davis, Marc Goadrich (2005). *The Relationship between Precision-Recall and ROC Curves.* University of Wisconsin Madison, USA

[30] Bob Nisbet, Gary Miner, John Elder, Thomas Hill, Dursun Delen, Andrew Fast (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications.* ISBN 978-0-12-386979-1

[31] Taboada, Maite, Julian Brooke, and Manfred Stede (2009). *Genre-based paragraph classification for sentiment analysis.* Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue, pages 62–70, London.

[32] Prof. Dr. Tobias Kollmann (2011) E-Business. Grundlagen elektronischer Geschäftsprozesse in der Net Economy. Gabler Verlag, ISBN-10: 3834924520

[33] Whittaker, S., Issacs, E., O'Day, V. (1997). Widening the net. Workshop report on the theory and practice of physical and network communities. SIGCHI Bulletin, 29(3), 27-30.

[34] Jones, Q. (1997). Virtual-communities, virtual-settlements and cyber-archaeology: A theoretical outline. Journal of Computer Mediated Communication, 3(3).

[35] Rheingold, H (1993). *The Virtual Community, Homesteading on the Electronic Frontier.* Reading, MA: Addison-Wesley Publishing.

[36] Wellman, B. (2000). *Changing connectivity: A future history of Y2.03K'.* Sociological Research online, vol. 4(no. 4).

[37] Jenny Preece & Diane Maloney-Krichmar (2003) *Online Communities: Focusing on sociability and usability.* Handbook of Human-Computer Interaction, Lawrence Erlbaum Associates Inc. Publishers. Mahwah: NJ. 596-620.

[38] Richard Millington (2010). *Different Types Of Communities.* http://www.feverbee.com/2010/11/different-types-of-communities.html (Sep 2012)

[39] Joshua Paul (2011). *5 Types of Online Community for Business.* http://info.socious.com/bid/48110/5-Types-of-Online-Community-for-Business-Part-2 (Sep 2012)

[40] Mark Kelly (2010). *Types of websites & online communities.* http://www.vceit.com/pages/online-community-websites.htm (Sep 2012)

[41] *Die Entwicklung des MMOG* http://www.online-spiele.me/die-entwicklung-des-mmog (Sep 2012)

[42] *World of Warcraft* http://de.wikipedia.org/wiki/World_of_Warcraft (Sep 2012)

[43] S. Sumathi,S.N. Sivanandam (2006). *Introduction to Data Mining and its Applications.* Springer XXII, 828 p. 108 illus.

[44] Myles Anderson (2013). *79% Of Consumers Trust Online Reviews As Much As Personal Recommendations.* searchengineland.com/2013-study-79-of-consumers-trust-online-reviews-as-much-as-personal-recommendations-164565 (Jan 2014)

[45] Alan Hall (2012). *To Succeed as an Entrepreneur, Know Your Customer.* http://www.forbes.com/sites/alanhall/2012/06/14/to-succeed-as-an-entrepreneur-know-your-customer/ (Sep 2012)

[46] Lucy Handley (2012). *It pays to know your target audience.* http://www.marketingweek.co.uk/opinion/it-pays-to-know-your-target-audience/4001836.article (Sep 2012)

[47] Dwi A.P. Rahayu, Shonali Krishnaswamy, Cyril Labbe, Osadi Alhakoon (2010). *Web Services for Analysing and Summarising Online Opinions and Reviews.* Springer-Verlag, LNCS 6481, pp 136-149

[48] David Harel, Yehuda Koren et al. (2000). *A Fast Multiscale Method for Drawing Large Graphs.* Springer, LNCS 1984, pp. 183-196

[49] Mourad Louha (2009). *Netzwerke visualisieren mit NodeXL.* `http://www.excel-ticker.de/netzwerke-visualisieren-mit-nodexl/` (Jan 2013)

[50] Ms. Anjali Ganesh Jivani (2011). *A Comparative Study of Stemming Algorithms. Int.* J. Comp. Tech. Appl, Vol 2 (6), ISBN:2229-6093

[51] Rosaria Silipo, Phil Winters, Killian Thiel, Tobias Kötter (2012). *Creating Usable Customer Intelligence from Social Media Data: Clustering the Social Community*. `https://knime.org/files/fillpdf/knime_social_media_data_clustering_whitepaper.pdf` (Nov 2012)

[52] Dr. Killian Thiel, Tobias Kötter, Dr. Michael Berthold, Dr. Rosaria Silipo, Phil Winters (2012). *Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining.* `http://knime.org/files/knime_social_media_white_paper.pdf` Revision 120403F (Nov 2012)

[53] Dr. Killian Thiel, Dr. Michael Berthold (2012). *The KNIME Text Processing Feature: An Introduction.* `http://www.knime.org/files/knime_text_processing_introduction_technical_report_120515.pdf` Revision 120403F (Nov 2012)

[54] Dr. Killian Thiel, Tobias Kötter, Dr. Michael Berthold, Dr. Rosaria Silipo, Phil Winters (2012). *Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining.* `http://knime.org/files/knime_social_media_white_paper.pdf` Revision 120403F (Nov 2012)

[55] Neil McGuigan (2013). *Vancouver Data Blog.* `vancouverdata.blogspot.com` (Apr 2013)

[56] Sheamus McGovern (2012). *Sentiment Analysis in Rapid Miner.* `http://www.corequant.com/?p=1` (Feb 2013)

[57] Victoria Pernik, Christian Schlögl (2006). *Möglichkeiten und Grenzen von Web Structure Mining am Beispiel von informationswissenschaftlichen Hochschulinstituten im deutschsprachigen Raum.* Diploma Thesis, Karl- Franzens- Universität Graz, `http://www.phil-fak.uni-duesseldorf.de/infowiss/content/forschung/publikationen/ChristianSchloegl_DGD.pdf` (Feb 2013)

[58] Johannes Fürnkranz (2002). *Web Structure Mining: Exploiting the Graph Structure of the World Wide Web.* American Journal of Applied Sciences 7 (6): 840-845, 210 ISSN 1546-9239