UPTEC STS08 004
Examensarbete 20 p
Februari 2008

A study of term weighting approaches for short-length documents

Jesper Zidén

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

A study of term weighting approaches for short-length documents

Jesper Zidén

This thesis involves a study of different approaches for term weighting as part of an ongoing development project at Ericsson Research. The purpose of this study is to determine which approach has the most suitable performance for retrieval of short-length documents. In text categorization, term weighting methods assign appropriate weights to the terms to improve the classification performance and retrieval of information. The aim of the thesis is to make an extensive comparative study of different term weighting approaches under controlled conditions. In this study an insightful analysis of the term discriminating power for the term weighting approaches is made from a qualitative perspective. The controlled experimental results showed that this newly proposed tf*rf approach is performing better than other widely used term weighting approaches. The contribution of the thesis is a recommendation of an effective term weighting method to enhance the performance of information retrieval from short length documents.

Handledare: Olof Lundström
Ämnesgranskare: Arne Andersson
Examinator: Elisabet Andrésdóttir
ISSN: 1650-8319, UPTEC STS08 004

Sammanfattning

När ett informationssystem ska utvecklas finns det många olika aspekter att ta hänsyn till för att de ska prestera så bra som möjligt. Forskningsavdelningen på Ericsson AB arbetar med ett pågående projekt för att utveckla en ny användartjänst. Detta examensarbete utgör en del av detta projekt och syftar till att studera olika tillvägagångssätt för termviktning av korta dokument.

Inom av informationsåtervinning av textdokument kan det ofta vara komplicerat att skilja relevanta dokument från irrelevanta dokument. Användandet av termviktnings metoder går ut på att kategorisera dokumenten för att göra det lättare hitta de dokument som verkligen är relevanta för användaren av informationssystemet. Så att systemets inhämtning av information sker på bästa möjliga sätt.

Det här arbetet syftar till att under kontrollerade experiment undersöka olika metoder för att utröna vilken metod som anses mest lämplig för att hitta korta dokument. Resultatet av arbetet visar att en ny metod för termviktning, tf*rf, presterar bättre än andra vanligen använda metoder. Bidraget från detta arbete är en rekommendation av vilken metod som är mest lämpad för informationsåtervinning av korta textdokument.

Acknowledgments

I would like to thank all the people who contributed to the realization of this Master's thesis report. First of all I would like to thank my colleague Joseph Vimal, for valuable collaborative work and support throughout the project.

I would like to thank my supervisor at Ericsson Olof Lundström, who always has been positive and inspiring. I also would like to thank my supervisor at Uppsala Universitet Professor Arne Andersson, for giving new input, correcting errors and giving valuable comments.

Other people I would like to thank are colleagues at Ericsson AB who gave me useful comments about the thesis research at the department of KI/EAB/TGE End-user mobile services and applications. They gave me valuable information and useful comments, which guided my work in a positive direction.

This thesis is dedicated to my always supportive and encouraging Anna.

Jesper Zidén
Stockholm, Dec 2007

Table of contents

Abbreviations

TF	term frequency
IDF	inverse document frequency
RF	relevance frequency

Definitions

- Short length document – Merely the Ads within the AdBoard system, containing approximately 1-40 word length passages.
  - Example: ”I want a bike”
- Term – the words used within Ads.
  - Example: I, want, a, bike
- Term weight – the importance weight given to a certain term.
  - Example: I=.2, want = .32,
- A match – when an Ad matches to a certain level of correspondence to another Ad inside the AdBoard system.
- Ad 1 : ”I lost a bike” Ad 2: ”I have found a bike”

1. Introduction

1.1 Background

In the information society of today, it is generally acknowledged within the field of computer science that computer users experience an excess of available information. With the increasing volume of obtainable information the users are not faced with the problem of shortage of data but instead they face the difficulty of finding relevant data. Providing effective search methods for relevant information becomes more and more essential in Information retrieval (IR) systems of today. An IR system assists the user to store, manipulate and retrieve useful data in form of a document.[1] IR systems are traditionally concerned with fulfilling a one-time information need. As a result the user has to be able to express what she is looking for in a straightforward and quick way, e.g. as a query of free language terms. Most versions of IR systems involve text search based methods that accept as input a query or limited-length textual phrase and produce as output a list of potentially relevant documents.[2]

The main duty of the IR system is to deliver accurate information that matches user's request efficiently and effectively. Since detecting relevant information depends on the performance of term weighting approaches and similarity measures, much research has been carried out to try improving and developing existing term weighting approaches further.[3]

In this paper, the author examines the results of applying Term Frequency Inverse Document Frequency (TF.IDF) to determine what words in a set of documents might be more favourable to use in a query. As the term implies, TF.IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF.IDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user. [4] This thesis provides evidence that this simple algorithm efficiently retrieve relevant documents.

1.1 Problem statement

The aim of this master's thesis is to conduct a study of three different term weighting approaches. The purpose of this study is to determine which approach has the most suitable performance for retrieval of short-length documents.

The author hopes that this study will illuminate new promising findings by focusing on comparing recently presented research findings and bringing valuable insight and inspiration to the development of IR systems.

---

[1] Kowalski, G. (1971)
[2] Nanas, N. Uren, V. and De Roeck A., (2003)
[3] Jones, K. S. (1972)
Salton, G. and Buckley, C., (1988)
Salton, G. and McGill, M., (1983)
[4] Ramos, J. (2003)

1.2 Assignment framework

This thesis work has been part of an ongoing project at Ericsson Research. The work of the "End-user mobile services and applications department" at Ericsson Research involves both developing and analyzing mobile services. A common problem with many new mobile services is that they do not reach a sufficient number of users. One possible explanation to this might be poor usability caused by services not matching the user's requests.

Depending on which applications and which term weighting methods that become dominant the architecture of mobile networks is affected in different ways. The reason for the appointment of this master thesis work was that "End-user mobile services and applications department" wanted to investigate if there could be angles in a application called *AdBoard* that would benefit from studying different term weighting approaches.

1.3 Limitations

Informally, query retrieval can be described as the task of searching a collection of data, be that text documents, databases, networks, etc., for specific instances of that data. First, I will limit myself and this thesis to searching only collections of English language documents. The second limitation of the thesis is that, it does not include scalability issues of the term weighting approaches; the reason for this focus is that these issues as this have been very well studied elsewhere.

1.5 Reading directions

Chapter 2, *An overview of term weighting* serves as an introduction and background to term weighting. Chapter 3, *Methodology*, describes the research process and chosen method. In Chapter 4 *Term weighting approaches* the different studied approaches are briefly defined, followed by the experimental setup in Chapter 5 *Term weighting experiments*. This chapter describes the experimental settings, performance measures and results.

Chapter 6, *Conclusions* summarizes the important aspects of the experiments. After the conclusions section follows a section of recommendations where solutions and further research areas are suggested. Lastly, the references are found in chapter 7, *References.*

2. An overview of term weighting

This section compromises an overview of the concept of term weighting followed by an introduction of the mathematical background of the approaches. Subsequently the algorithm as it has been implemented in java code will be briefly presented. The section then stresses some important features and issues concerning term weighting like content representation and how the determination of the term weights allows differentiation of distinguishing the important terms from the less crucial.

2.1 Background to TF.IDF

In 1972, Karen Spärck Jones published in the Journal of Documentation a paper called "A statistical interpretation of term specificity and its application in retrieval". The measure of term specificity first proposed in that paper later became known as inverse document frequency, or IDF; it is based on counting the number of documents in the collection being searched which contain (or are indexed by) the term in question. The intuition was that a query term which occurs in many documents is not a good discriminator, and should be given less weight than one which occurs in few documents, and the measure was an heuristic implementation of this intuition.[5]

The intuition and the measure associated with it, proved to be a giant leap in the field of information retrieval. Coupled with TF (the frequency of the term in the document itself, in this case, the more the better), it found its way into almost every term weighting scheme.
The class of weighting schemes known generically as TF.IDF, which involve multiplying the IDF measure (possibly one of a number of variants) by a TF measure (again possibly one of a number of variants, not just the raw count) have proved extraordinarily robust and difficult to beat, even by much more carefully worked out models and theories. It has even made its way outside of text retrieval into methods for retrieval of other media, and into language processing techniques for other purposes.[6]

The method examined with more detail is Term Frequency Inverse Document Frequency (TF.IDF). This weighing approach can be categorized as a statistical procedure, though its immediate results are deterministic in nature. Though TF.IDF is a relatively old weighing scheme, it is simple and effective, making it a popular starting point for other, more recent algorithms.[7]

2.2 Mathematical Framework

This subsection will give a quick informal explanation of TF.IDF before proceeding. Essentially, TF.IDF works by determining the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire document set. Intuitively, this calculation determines how relevant a given word is in a particular document. Words that are common in a single or a small group of documents tend to have higher TF.IDF numbers than common words

[5] Jones, K. S., (1972).
[6] Harman, D., (2006)
[7] Salton, G. and Buckley, C. (1988)

such as articles and prepositions. The formal procedure for implementing TF.IDF has some minor differences over all its applications, but the overall approach works as follows. Given a document collection D, a word w, and an individual document $d \in D$, one can calculate

$$wd = fw, d * \log (|D|/fw, D) \quad (1)$$

Where fw, d equals the number of times w appears in d, $|D|$ is the size of the corpus, and fw, D equals the number of documents in which w appears in D.[8]

There are a few different situations that can occur here for each word, depending on the values of fw, d, $|D|$, and fw, D, the most prominent of which I will examine. Assume that $|D| \sim fw, D$, i.e. the size of the set is approximately equal to the frequency of w over D. If $1 < \log (|D|/ fw, D) < c$ for some very small constant c, then wd will be smaller than fw, d but still positive. This implies that w is relatively common over the entire set but still holds some importance throughout D. For example, this could be the case if TF.IDF would examine the word Jesus over the New Testament. This is also the case for extremely common words such as articles, pronouns, and prepositions, which by themselves hold no relevant meaning in a query (unless the user explicitly wants documents containing such common words). Such common words thus receive a very low TF.IDF score, rendering them essentially negligible in the search. Finally, suppose fw, d is large and fw, D is small. Then $\log (|D|/ fw, D)$ will be rather large, and so wd will likewise be large. This is the case which is most interesting, since words with high wd imply that w is an important word in d but not common in D. This w term is said to have a large discriminatory power. Therefore, when a query contains this w, returning a document d where wd is large will very likely satisfy the user.

2.3 Encoding TF.IDF

The code for TF.IDF is elegant in its simplicity. Given a query q composed of a set of words wi, I calculate wi, d for each wi for every document $d \in D$. In the simplest way, this can be done by running through the document collection and keeping a running sum of fw, d and fw, D. Once done, one can easily calculate $wi\ d$ according to the mathematical framework presented before.[9] Once all wi, d.s are found, we return a set D^* containing documents d such that the following equation is maximized:

$$\Sigma i\ wi, d \quad (2)$$

Either the user or the system can arbitrarily determine the size of D^* prior to initiating the query. Also, documents are returned in a decreasing order according to equation (2). This is the traditional method of implementing TF.IDF.[10] The thesis will discuss extensions of this algorithm in later sections, along with an analysis of TF.IDF according to the experimental results.

[8] Salton, G. and Buckley, C. (1988)
[9] Witten, I., Moffat, A. and Bell, T. (1999)
[10] Goodrich, T. and Tmassia, R. (2001)

2.4 Term weighting issues

Term weighting is faced by two main questions. First, what appropriate terms are to be included in the content representation of documents? Second, is the determination of the term weights capable of distinguishing the important terms from the less crucial? These two questions will be answered in the following sections

2.4.1 Content representation

The content representation that dominates the text classification literature is known as the "bag of words". This is a low level of content representation. For most bag of words representations, each feature corresponds to a single word found in the training set, usually with case information and punctuation removed. Often infrequent and frequent words are removed from the original text. Sometimes a list of stop words (functional or connective words that are assumed to have no information content) is also removed.[11]

Sometimes this word set is used with no further processing but more typically, there is some attempt to make the features more statistically independent. The most common way achieve independence is to remove suffixes from words using a stemming algorithm such as the one developed by Lovins.[12] Stemming has the effect of mapping several morphological forms of words to a common feature. For example the words *learner*, *learning*, and *learned* would all map to the common root *learn*, and this latter string would be placed in the feature set rather than the former three.[13] Stemming has not been considered as part of the thesis scope, mainly because its high level of complexity made it difficult to implement in the encoding of the approaches.

While a large number of document representations have been proposed, most of them use the same starting point, namely the words appearing in a document. In fact, a common choice is to represent a document as a "bag of words", i.e. a document is represented by the set of words appearing in it[14]. Another commonly used and slightly richer representation takes account of the frequency with which words appear in a specific document. Such representations ignore important aspects of a document, for instance the order in which words appear in the document, the syntax etc. Richer representations have also been proposed but the emphasis on words is usually retained.[15]

In this thesis I will present experiments to compare document representations which utilize only words and not word meanings. This is a high level of content representation. It is a characteristic of natural languages that the same word may assume different meanings in different contexts. For example the word "car" and "automobile" share the same meaning; the word "crane" may mean either a bird or a machine that lifts and moves heavy objects; and so on. The extra information received from word meanings has not proven to be fruitful enough to mirror the increased complexity.[16] Therefore the word level representation is used in this thesis.

[11] Harman, D., (2006)
[12] Lovins J. B. (1968)
[13] Scott, S. and Matwin, S., (1999)
[14] Manning, C.D. and Schuetze, H., (1999)
[15] D. Mladenic, (1998)
[16] Kehaigas, A. (2001)

2.4.2 Term weights

The main function of a term-weighting system is the enhancement of retrieval efficiency. Effective retrieval depends on two main factors: one, items likely to be relevant to the user's needs must be retrieved; two, items likely to be extraneous must be rejected. Two measures are normally used to assess the ability of a system to retrieve the relevant and reject the non relevant items of a collection, known as *recall* and *precision,* respectively. Recall is the proportion of relevant items retrieved, measured by the ratio of the number of relevant retrieved items to the total number of relevant items in the collection; precision, on the other hand, is the proportion of retrieved items that are relevant, measured by the ratio of the number of relevant retrieved items to the total number of retrieved items.[17]

In principle, a system is preferred that produces both high recall by retrieving everything that is relevant, and also high precision by rejecting all items that are extraneous. The recall function of retrieval appears to be best served by using broad, high-frequency terms that occurs in many documents of the collection. Such terms may be expected to pull out many documents, including many of the relevant documents. The precision factor, however, may be best served by using narrow, highly specific terms that are capable of isolating the few relevant items from the mass of non relevant ones. In practice, compromises are normally made by using terms that are broad enough to achieve a reasonable recall level without at the same time producing unreasonably low precision. The differing recall and precision requirements favor the use of composite term weighting factors that contain both recall- and precision-enhancing components.[18]

Three main considerations appear important in this thesis. First, terms which are frequently mentioned in individual documents, appear to be useful as recall enhancing devices. This suggests that a *term frequency* (TF) factor be used as part of the term-weighting system measuring the frequency of occurrence of the terms in the document or query texts.[19]

Second, term frequency factors alone cannot ensure acceptable retrieval performance. Specifically, when the high frequency terms are not concentrated in a few particular documents, but instead are prevalent in the whole collection, all documents tend to be retrieved, and this affects the search precision. Hence a new collection-dependent factor must be introduced that favors terms concentrated in a few documents of a collection. The well-known *inverse document frequency* (IDF) factor performs this function. The idf factor varies inversely with the number of documents *d* to which a term is assigned in a set of D documents. A typical idf factor may be computed as log *d/D.*[20]

Term discrimination considerations suggest that the best terms for document content identification are those able to distinguish certain individual documents from the remainder of the collection. This implies that the best terms should have high term frequencies but low overall

---

[17] Salton, G. and Buckley, C. (1988)
[18] Ibid.
[19] Lan M., Tan C-L., Low H-B (2006)
[20] Jones, K. S., (1972).

collection frequencies. A reasonable measure of term importance may then be obtained by using the product of the term frequency and the inverse document frequency (TF.IDF).[21]
The term discrimination model has been criticized because it does not exhibit well substantiated theoretical properties. This is in contrast with the probabilistic model of information retrieval where the relevance properties of the documents are taken into account, and a theoretically valid *relevance frequency* (RF) weight is derived. The relevance frequency term weighting approaches weight defined as the proportion of relevant documents in which a term occurs divided by the proportion of non relevant items in which the term occurs is, however, not immediately computable without knowledge of the occurrence properties of the terms in the relevant and non relevant parts of the document set. A number of methods have been proposed for estimating the term relevance factor in the absence of complete relevance information, and these have shown that under well defined conditions the term relevance can be reduced to an inverse document frequency factor. The composite (TF.IDF) term weighting system is thus directly relatable to other theoretically attractive retrieval models.[22]

A third term-weighting factor, in addition to the term frequency and the inverse document frequency, appears useful in systems with widely varying vector lengths. In many situations, short documents tend to be represented by short-term vectors, whereas much larger-term sets are assigned to the longer documents. When a large number of terms are used for document representation, the chance of term matches between queries and documents is high, and hence the larger documents have a better chance of being retrieved than the short ones. Normally, all relevant documents should be treated as equally important for retrieval purposes. This suggests that a normalization factor be incorporated into the term weighting formula to equalize the length of the documents.[23]

2.5 Summary

These constraints are motivated by the following observations on some common characteristics of typical retrieval formulas. First, most retrieval methods assume a “bag of words” (more precisely, “bag of terms”) representation of both documents and queries. Second, a highly effective retrieval function typically involves a TF part, an IDF part, and a document length normalization part. The TF part intends to give a higher score to a document that has more occurrences of a query term, while the IDF part is to penalize words that are popular in the whole collection. The document length normalization is to avoid favoring long documents; long documents generally have more chances to match a query term simply because they contain more words. Finally, different retrieval formulas do differ in their way of combining all these factors, even though their empirical performances may be similar.

[21] Lan M., Tan C-L., Low H-B (2006)
[22] Salton, G. and Buckley, C. (1988)
[23] Nanas, N. Uren, V. and De Roeck A., (2003)

3. Methodology

This section explains the method by which this study was conducted. Each phase of the selected methodology is presented and discussed for the development of this thesis. At the end some considerations regarding the strengths and weaknesses of the chosen method is discussed.

In any type of scientific research there are several basic steps, which need to be followed in order to secure the completion of its principle. In this chapter, a series of steps was described in order to collect the data necessary to obtain answers to the research question stated in this thesis. These methodology issues are presented below in figure 1, which show a graphical overview of the methodology utilized in this thesis.

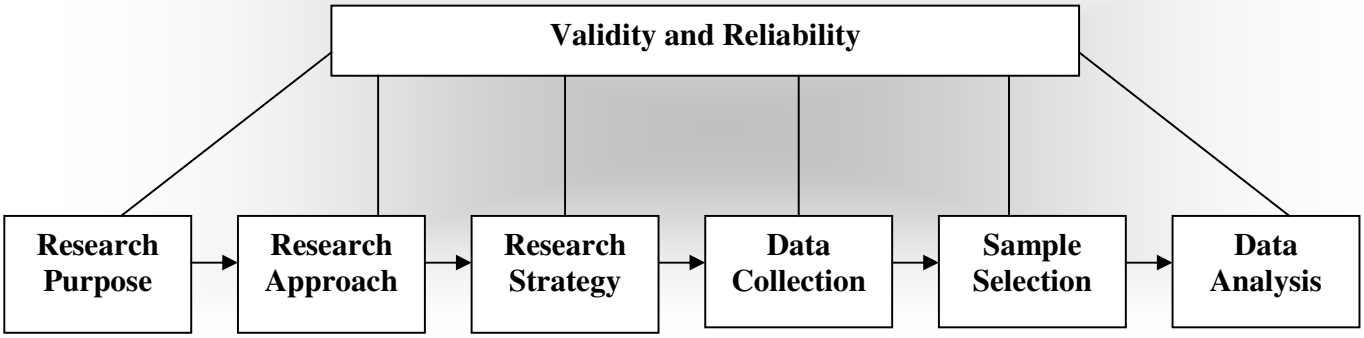

Figure 1: Schematic presentation of the Methodology (Foster)[24]

Although in a real case research it is not likely to resemble a linear and highly structured process like this. The research work has actually more been part of an ongoing and iterative process, switching back and forth between the different phases. Nonetheless, a visual model of the main phases involved in research can be really helpful to illustrate the work done to complete this thesis and illuminating the importance of validity and reliability when conducting scientific research.

3.1 Research purpose

The starting point of this was the identification of an area of interest, a purpose for conducting the research. The thesis was started as an assignment from a department at Ericsson Research to investigate if there could be angles in a concept called *AdBoard* that would benefit from a comparative study of different term weighting approaches. This research field has become more and more focused on by computer scientists and as part of that term weighting used in applications has also become more imperative.

3.2 Research approach

Regarding the problem statement, as described in section 1.2, it resulted from other interesting research questions that where touched upon through making a literature study of current research topics concerning term weighting. The main point of interest was that none of the current efforts

[24] Foster, T. (1998)

made within this research field was focused on the term weighting of short-length documents. Simply there was a gap in the knowledge of such studies, which made the importance of conducting such a study more understandable as well as there was practical need for the application development of *AdBoard*. In addition the decision also depended on what was practically feasible to carry trough within the scope of master's thesis.

3.3 Research strategy

There are two main methods to be considered when conducting research the qualitative and the quantitative one. Quantitative methods are applicable on research regarding measurable experiments and tangible results; the method is suitable on studies with the aim of comparing different approaches. [25]

3.4 Data collection

The first step was to review articles, books and web sources to gain knowledge from prior literature on the related topic. This resulted in the background presentation (see section 2) for this thesis. The next step was to construct the experiments, will be presented further in section 5 *Term weighting experiments*. The following two limitations of the thesis should however be noted, first neither the thesis or the experiments are focused on scalability issues of the term weighting approaches since this has been well studied elsewhere within this research field and falls out of the thesis scope. Second the thesis is limited to searching text documents written in English language.

3.5 Sample selection

The sample selection for the controlled experiments conducted as part of this thesis work was made in an effort to achieve a comparative measurement of the different term weighting approaches efficiency on short-length documents. But they will not be included as part of an appendix or such because corporate regulations state that the sample ads may not be published in the thesis as this is considered as potentially sensitive information by the thesis assignors

3.6 Data analysis

After having conducted the experiments followed by trying to interpret the experimental findings in section 6 *Result* it's important to keep in mind that the findings from the experiments and the results should be considered in regard of the assumptions made in the thesis.

3.7 Discussion of the method

In any academic thesis, the validity and reliability of drawn conclusions should be questioned for the selected method. With this selected method one should remember that the results are not statistically verified. For example the experimental findings in this thesis are made based on the controlled experiments and it would be interesting to see if the same performance can be observed on a wider and more general setting. Thus the conclusions made within the thesis (see section 7) should not be generalized without consideration.

[25] Andersen, I, (1998)

4. Term weighting approaches

This section will give a more detailed description of the studied term weighting approaches. The intention of the section is to shed light on what differentiates the implementation made of the term weighting approaches into the AdBoard service concept from a general information retrieval system. As well as stating the assumption on which the experimental design setup builds on.

4.1 The AdBoard concept

A brief description of the service concept is needed in order to understand how the term weighting approaches have been implemented within this thesis study. Due to corporate regulations only a short description of will be given. The aim is purely to give the reader a general knowledge about the service in order to be able to understand the results from the studies.

In this research a prototype bulletin board application for user posted advertisements called *AdBoard* was used. This prototype has been developed by Ericsson Research, it is important to note that *AdBoard* is not an existing, or yet even a future, product. The description of the service concept used for this thesis is the following. AdBoard is an interactive online bulletin board for all needs in everyday life with an intelligent matching engine and search of advertisements (ads) in a community**.** The service can be used from your mobile phone and/or your PC. Users post ads expressing their needs, wishes & wants in text, or they can search for ads created by other users. Ads are matched with each other by an intelligent matching engine and if a matching ad is found, the users get instant notifications by SMS/email.

4.2 Term frequency weighting (TF)

In text categorization, term weighting methods assign appropriate weights to the terms to improve the classification performance of information retrieval systems. In this study, the aim is to propose an effective term weighting approach to be implemented for retrieval of short-length documents. Specifically a recommendation that is feasible enough to be put into practice in the *AdBoard* service.

The traditional term weighting methods for text categorization are usually borrowed from information retrieval field. Here the simplest approach is using the raw term occurrence alone without other factors. Translated into the implementation of *AdBoard* the term frequency is defined to be:

TF = number of times a term occurs in an ad.

4.3 Term frequency * inverse document frequency weighting (TF.IDF)

The most popular term weighting approach is TF.IDF which has been widely used in information retrieval and has consequently been adopted by researchers in text categorization.[26] As noted in section 2.1 the famous TF.IDF approach has a long history and still it has been very hard to find

[26] Salton, G. and Buckley, C. (1988)

an approach that outperforms it. This approach is the one which is currently implemented in *AdBoard*. The term frequency * inverse document frequency weighting approach is defined to be:

TF.IDF = number of times a term occurs in an ad * by the total number of all ads divided by the number of ads containing the certain term, and then taking the 2 based logarithm of that quotient, see equation (1).

4.4 Term frequency * relevance frequency weighting (TF.RF)

Lastly the third term weighting approach studied in this thesis is a newly developed improvement of TF.IDF, proposed by Lan *et al*.[27] The term frequency * relevance frequency approach have been introduced to improve the discriminating power of terms in the traditional information retrieval field as well as in text categorization. In section 2.4.2 it was stated that t*erm discrimination* considerations suggest that the best terms for document content identification are those able to distinguish certain individual documents from the remainder of the collection. This has been the target issue at hand for the development of the *AdBoard* service and an improvement of the retrieval of short-length documents is the whole thesis. The implementation the relevance properties of the documents are taken into account and a theoretically valid *relevance frequency* (RF) weight is derived in this approach. Lan *et al* have assigned the constant value 2 in the RF formula because the base of this logarithmic operation is 2.[28] Thus the term frequency * relevance frequency is defined as:

TF.RF = number of times a term occurs in an ad * the proportion of relevant documents in which a term occurs divided by the proportion of non relevant items in which the term occurs.

4.5 Experimental assumptions

However, it is imperative to note that RF is not immediately computable without knowledge of the occurrence properties of the terms in the relevant and non relevant parts of the document set. Therefore to be able to implement this approach one must first as it is customary to do in text categorization is to create a form of *supervised* learning with the purpose of makes use of prior information. In this study, one can classify the term weighting methods into two categories according to whether the method makes use of this known information on the membership of training documents, namely, supervised term weighting method and unsupervised term weighting method. For instance, the traditional methods borrowed from information retrieval field, such as TF and TF.IDF and its variants, belong to the unsupervised term weighting methods.[29] And as stated above TF.RF belongs to the supervised term weighting methods. Since the TF.RF approach is regarded as an enhanced method making use of prior knowledge it should perform better then the other approaches studied within this thesis. So the following hypothesis is made:

The key hypothesis in the thesis is that the experimental results should show a premier performance from the improved approach (TF.RF).

---

[27] Lan M., Tan C-L., Low H-B (2006)
[28] *ibid*
[29] Salton, G. and Buckley, C. (1988)

5. Term weighting experiments

This section describes the experimental settings, how to evaluate retrieval efficiency & performance measures and finally the experimental results.

5.1 Experimental settings

So as to investigating how do these three term weighting approaches behave I have designed the following methodology for the experiment.

- First create a set of fixed 100 sample ads to measure performance of the different approaches. Followed by an encoding and implementation of each of the different term weighting approaches in java code.

- Second the experiment tries to answer the following question:
  How does this matching approach behave?
  - Use the set of hypothetical sample ads consisting of terms and their weights.
  - Create some hypothetical user ads.
  - How are the ads ranked, depending on the weights of their terms?

- This was studied by repeatedly test running the sample ads to evaluate their general performance of the approaches on both a micro and macro scale. Since it is a controlled performance comparison the number of pre assumed matching ads are known. Which is the main concern for the supervised term weighting TF.RF approach.

The entire set of sample ads have been put through a document length normalization step to avoid favoring long documents; long documents generally have more chances to match a query term simply because they contain more words. Normally, all relevant documents should be treated as equally important for retrieval purposes. [30] This suggests that a normalization factor be incorporated into the term weighting formula to equalize the length of the documents as stated in section 2.4.2

5.2 Evaluating retrieval efficiency

In order to compare the term weighting approaches, we need some way of quantifying their performance. The approaches performance should be based on the total term weighting it imposes on the sample set with respect to a test run. A number of different methods have been suggested for this. None are entirely satisfactory, but this is a natural consequence of attempting to represent multidimensional behavior with a single representative value. Below, the thesis will define two important measures of efficiency: recall and precision. But to help illustrate the term weighting approaches efficiency a two way contingency table will be used. The table contains four cells:

[30] Nanas, N. Uren, V. and De Roeck A., (2003)

- a counts the # Ads assigned as matching and correct,
- b counts the # Ads assigned as matching but incorrect,
- c counts the # Ads assigned as not matching but incorrect,
- d counts the # Ads assigned as not matching and correct.

	Matching is correct	**Not matching is correct**
Assigned as matching	**a**	**b**
Assigned as not matching	**c**	**d**

Table 1 A contingency table (adapted from Yang)[31]

5.2.1 Recall and precision

Classification efficiency and retrieval performance are usually measured by using *precision* (*p*) and *recall* (*r*). *Precision* is the proportion of truly positive examples labelled positive by the system that were truly positive and *recall* is the proportion of truly positive examples that were labelled positive by the system. This leads to the following definitions:

The precision P and recall R of a ranking method is defined to be

$$P = \frac{\#\text{Ads found as matching and correct}}{\text{total \#Ads found}} = \frac{a}{a+b} \quad \text{(if } a+b>0\text{, otherwise P=1)}$$

$$R = \frac{\#\text{Ads found as matching and correct}}{\text{total \#Ads correct}} = \frac{a}{a+c} \quad \text{(if } a+c>0\text{, otherwise R=1)}$$

5.2.2 Performance Evaluation

The F1 function which combines *precision* and *recall* is computed as:

$$F1 = (2pr) / (p+r)$$

Usually, F1 function is estimated from two ways, i.e. micro-averaged and macro-averaged. The two measures may give quite different results, that is, the ability of a classifier to behave well also on categories with low generality (i.e., categories with few positive training instances, that is very rare terms) will be emphasized by macro-averaged and much less so by micro-averaged. A commonly used measure in term weighting method comparison is the break-even point of recall and precision i.e. when *r* and *p* are tuned to equal. Often the break-even point is close to the optimal score of F1.[32]

[31] Yang 1997

[32] Lewis ,D. (1992) Yang, Y (1999)

5.3 Experimental results

Table 2 summarizes the results of all the three term weighting approaches investigated in this study. The experimental results origin from test running the set of fixed 100 sample ads repeatedly and averaging the results on both a micro and macro scale. The measures range from 0-1 (0-100%). The trend is distinctive that TF.RF is outperforming the other approaches in regard to recall and precision, but the margin between them is some what limited. The best break-even points are achieved by TF.RF as well, but yet again the difference between approaches is diminutive. The results which should be given the most importance are the micro- and macro averaged break-even point since this performance measurement illuminates the different approaches performance on general respectively on specific rare term weighting.

Approach	**Micro - Recall**	**Micro - Precision**	**Micro - Breakeven point**	**Macro - Breakeven point**
TF	**0.6565**	**0.6566**	**0.6572**	**0.6335**
TF.IDF	**0.6567**	**0.6588**	**0.6578**	**0.6407**
TF.RF	**0.6810**	**0.6800**	**0.6805**	**0.6604**

Table 2 Experimental results

The results were produced as a list of potentially matching ads by the different approaches implemented as algorithms in Java coding. The recall and precision of a category ranking of matches and non matches is similar to the corresponding measures used in text categorization and information retrieval. The documents are returned in a decreasing order according to equation (2) and the top documents correspond to the best matching ads. Given the document as the input to the algorithm, and a ranked list of matching documents as the output, the recall and precision at a particular threshold on the ranked list are defined to be either above the threshold i.e. regarded as matches or successively if they are below the threshold i.e. regarded as non matches. As stated in section 2.3 the encoding of TF.IDF either the user or the system can arbitrarily determine the size of D^* prior to initiating the query. Meaning the user can decide for himself how high or low the threshold should be, that is to say depending on his or her demands how specific the matching short length documents need to be. It is readily observed that there is a balance between getting every matching document but also the possibility of getting an excess of non matching irrelevant documents. This is up to the user to decide the desired specifications.

6. Conclusions

This section aims to analyze the results presented in chapter five of this thesis. It discusses the validity and reliability of the conducted study. Alternative possible methods as well as suggested further research are also discussed.

6.1 Discussion of Findings

The results that were presented in section 5.3 are interesting in different ways. First the representativeity and validity of the experiment is discussed in detail based on the experimental assumption made in section 4.5. For the reason that it is an improvement of the other approaches and each term should consequently be assigned more appropriate weights and therefore giving more relevant matching.

I would like to point out that the observations above are made based on controlled experiments, therefore somewhat limiting the reliability of the results. It would be interesting to see if the AdBoard service concept can observe similar results on a wider and more general setting. The reliability of the research indicates how well it produces the same results on separate occasions. It should not matter who conducts a study, another evaluator who follows the same procedure should get similar results. Higher reliability can be achieved through a more structured and standardized methods and data gathering techniques. This is somewhat limited by the fact of due to corporate regulations the sample ads may not be published in the thesis as this is considered as potentially sensitive information by the thesis assignors. Nonetheless the validity of the conducted experiment is strong thanks to the carefully designed experiments and repeatedly running them to make sure the result trends are real. The validity is concerned with whether an evaluation technique measures what it is supposed to measure and this thesis assignment fulfills that goal.

6.2 Recommendations

Based on the experimental findings of the thesis, based on the key hypothesis in the thesis is that the experimental results should show a premier performance from the improved approach (TF.RF). The thesis results suggest that TF.RF be used as term weighting approach for retrieval of short-length documents as it performed consistently well on both micro and macro scale.

Another option, since the difference in performance was not immense, would be using the most popular method TF.IDF. This is currently implemented in AdBoard with the possible extension of further performance increasing possibilities. See future research below.

6.3 Future research

This section lists possible future research suggestions to further determine which approach has the most suitable performance for retrieving short length documents. When evaluating the potential improvements of the work completed within this thesis one suggestion would be extending term weighting approaches by applying term weighting methods to other text-related

applications. The comparison between approaches could instead be focusing on certain language issues and semantics.

Another suggestion would be to try including user relevance feedback in the term weighting. Information retrieval is traditionally concerned with satisfying a one time information need. As a result the user has to be able to express what he or she is looking for in a straightforward manner that is as a query of free language terms. The query is the only information initially available about what the user is looking for. Weighting of query terms can be accomplished if additional relevance information is available. In information retrieval such information can be acquired by user feedback for the documents retrieved so far.

6.4 Conclusions

All the aims of the thesis presented in section 1.2 problem statement have been reached. This section contains a summary of them together with the descriptions of how they were reached.

The aim of this master's thesis is to conduct a study of three different term weighting approaches. The purpose of this study is to determine which approach has the most suitable performance for retrieval of short-length documents.

The study of which term weighting approach has the most suitable performance have been detailed in the recommendation section of this chapter and as stated above the TF.RF is most suitable for this goal. But possible extended research suggestions also have a promising future.

This thesis study has also illuminated new promising findings by focusing on a new feature of information retrieval namely the AdBoard service concept. By comparing the recently presented TF.RF approach and the relatively old TF.IDF approach it has been proven that this solution, simple as it may be, it is still effective and hard to beat. This is clearly shown by the small difference between the three studied approaches. Hopefully this thesis will contribute with valuable insight and inspiration to the development of other information retrieval systems in the future.

7. References

7.1 Articles

Andersen, I., (1998) *Den uppenbara verkligenheten: val av samhällsvetenskaplig metod*, Studentlitteratur Lund

Harman, D., (2006) *The history of idf and its influences on IR and other fields,* p. 69-98

Foster, T., (1998) *Industrial market communication: an empirical investigation on the use of marketing communication tools,* Luleå

Jones, K. S. (1972) *A statistical interpretation of term specificity and its application in retrieval*, Journal of documentation, 28(1): 11-20.

Kehaigas, A. (2001) *A Comparison of Word- and Sense-based Text Categorization Using Several Classification Algorithms* Springer Netherlands p.227-247

Kowalski, G. (1997) *Information retrieval systems – theory and implementation*, Kluwer Academic Publishers, Boston, Ma.

Nanas, N. Uren, V. and De Roeck, A. (2003) *A comparative study of term weighting methods for information filtering* 4th International Workshop on Natural Language and Information Systems, p. 13-17

Ramos, J. (2003) *Using TF-IDF to Determine Word Relevance in Document Queries* Rutgers University

Salton, G. and Buckley, C. (1988) *Term-weighting approaches in automatic text retrieval. Information processing and management*, 24(5): 513-523.

Salton, G. and McGill, M. (1983) *Introduction to modern information retrieval,* McGraw Hill, New York

Scott, S. and Matwin, S. (1999) *Feature engineering for text classification* Proceedings of 16th International Conference on Machine Learning

Yang, Y., (1997) *An evaluation of statistical approaches to text categorization* Inf. Retr. (1-2):69-69

7.2 Books

Witten, I., Moffat, A. and Bell, T. (1999) *Managing gigabytes* 2nd edition Morgan Kaufmann New York
Goodrich, T. and Tmassia, R. (2001) *Algorithm design* Wiley New York