Information Retrieval models based on electrical circuit analysis using Ohm's law

Bachelor's Research

V.L. van Donselaar

v.l.vandonselaar@student.utwente.nl

Abstract

The TF×IDF weighting function is a well-known and proven model in modern information retrieval systems. The model allows documents to be ranked by relevance based on the frequency of terms issued by a user's query. Despite the fact that it yields good results, an clarification for its success is not so obvious. Attempts have been made to explain the model in terms of statistics or common sense. This paper tries to find similarities and differences with the theory of network analysis. A simplified network model based on the principle of an electrical circuit acts as a guide to gain understanding of the model's operation. The correctness of this model is tested by implementing it as a function of the Terrier Information Retrieval System, whereupon it is evaluated against Terrier's predefined TF×IDF model. Results show that the precision of the network model is not as high as a TF×IDF model would typically achieve. Nonetheless the network model shows a new approach for calculation of the document score based on multiple termed queries, which improves the precision of the top 10 results.

Keywords

Information Retrieval, Inference Networks, TF-IDF, Ohm's law

1 Introduction

Information retrieval is a field of research which exists for multiple decades. Over the years, several retrieval models have been invented, allowing the users of an information retrieval system to query for relevant documents [Singhal, 2001]. The omnipresence of search engines like Google, Yahoo and Bing are proof of the definitive success of retrieval systems. The definition of an information retrieval system is very broad. A system based on just a simple matching model does not necessarily take responsibility for the *ranking* of results: the first document retrieved is not necessarily the most relevant. This is not problematic as long the result sets keep small. With data volumes growing over time, ranking models became an important aspect in the field. A user will not take time to evaluate thousands of documents returned by a system: the results must be ordered by relevance. Key in ranking functions is term weighting [Salton and Buckley, 1988]. Term weighting is a technique which tries to determine which terms are important and which are not. A document containing a few important words might therefore be correctly designated as being more relevant than a document containing more, but unimportant words. A well-established model among term weighting is the term-frequency × inverse document frequency (TF×IDF) term-weighting function. This ranking model (which sums up the scores of individual terms) is a result of years of reasoning and trail-and-error. Although offering quite remarkable results, an in-depth understanding or explanation of its success is not trivial. Over time, insightful attempts have been made to understand the success of this model in terms of statistics [Hiemstra, 2000, Robertson, 2004]. This paper interprets the TF×IDF model in terms of electrics and circuit analysis. Based on this interpretation, a simplified model based on Ohm's law is evaluated against TF×IDF. Electric circuits can be analyzed by using network models. The concept of network models is however not a new approach to information retrieval. Turtle and Croft [1990, 1991] already closed the gap between the worlds of information retrieval and network analysis. Nonetheless the perspective of their work, which is based on inference networks, differs from a typical electrical circuit analysis. Ohm's law is chosen as a starting point, ignorant of previous work based on Bayesian networks. Hopefully this will provoke new insights in ranking model research.

1.1 Contents

In the following section, a short overview of ranking models is given. Section 3 elaborates on TF×IDF (and its variants) more in-depth. Thereafter, section 4 describes the commonalities and differences found with the analogy of circuit analysis. The section is meant to create a bridge to the next section, while clarifying the origin of the idea to express term frequencies and relevance in terms of resistance and potential. Section 5 shows how the circuit approach would work, on which section 6 draws its conclusions.

2 Retrieval and ranking models: an overview

The standard boolean model of information retrieval, which dates back until the 1960's, is a very simple approach to determine whether a document *could* be relevant or not and is therefore still contemporary [Manning et al., 2008]. Unfortunately, this model is

only appropriate to tell whether a term occurs within a document. Documents are usually returned in natural or random order. Already in early history, attempts have been made to improve the standard boolean model, in order to add better ranking / term weighting capabilities. As such, the Vector-space model emerged [Salton et al., 1975], and other proposals followed [Salton et al., 1983]. This is a common trajectory of research. Starting with a simplified model allows step-by-step improvement. In electrical circuit analysis, this is a very common practice. To save time and effort, models are always simplified as long the net results lies within acceptable boundaries. If not, the model is refined. The added complexity hopefully improves the model's precision, like it is in information retrieval.

2.1 Types of ranking models

Ranking models have been researched over time to improve the order in which retrieval systems return their results. Since these models are not trivial at all [Hiemstra, 2001], large varieties have been developed. They can roughly be categorized as follows.

Constant ranking
: This type of ranking model applies the same score to all documents. The standard boolean model is a trivial example of such model.

Term frequency
: Term frequency (and optionally additional inverse document frequency) involve counting the number of term occurrences in documents. Improved variants like BM25 are also part of this category.

Machine learning
: Machine learned ranking (MLR) is a method of which a ranking model is constructed based on training data, feeded to machine learning algorithms. First records on MLR were published by Fuhr [1989]. Its operations falls beyond the scope of this paper.

Static ranking
: Static rankings are fully independent from queries. Boolean matching determines the result set, which is ordered by a precalculated score. This ranking method is primarily well suited for web-scale search engines because it reduces the processing power required per query.

3 The TF-IDF ranking model

The TF×IDF function calculates the weight of a single term *(t)*, given a set of documents *(d)* in which it occurs as a part of a superset *(D)* which is the entire corpus. TF×IDF is the product of the term frequency *(TF)* and the inverse document frequency *(IDF)*:

$$\text{tfidf}(t,d,D) = \text{tf}(t,d) \times \text{idf}(t,D)$$

Both tf and idf exist in different varieties [Manning et al., 2008]. Common expressions for the term frequency $\text{tf}(t,d)$ include:

Boolean
: The term frequency is 1 iff the term occurs in d. Otherwise it is 0.

Natural
: The actual number of times the term occurs in document d.

Logarithm
: The term frequency is determined by $\text{tf}(t,d) = 1 + \log(f(t,d))$. If the term does not occur, the term frequency is 0.

Normalized
: Often, the natural term frequency is normalized to prevent long documents of getting relatively high term frequencies.

For the inverse document frequency, a constant value can be chosen. This will result in just using the tf term, rather than tf×idf. Usually the idf function is defined as:

$$\text{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Herein is $|D|$ the total number of documents. A high score is given to a term which occurs often in one particular document while being sparse in other documents.

4 TF-IDF using a physical approach

Research done by Turtle and Croft [1990, 1991] continues on the idea of expressing documents and search terms as an inference network. Their work proposes a model in which they estimate the probability that a users information need (a query) is met given a document as 'evidence'. The estimation is based on properties of the network, which consists of document nodes, text representation nodes and concept representation nodes. The query, which is also a network, is analyzed based on the links between both networks. Terms and concepts corresponding to a query are modeled as a vertex between the networks nodes. Without going further into the details of the networks properties, the research makes clear that certain characteristics of the TF×IDF can be explained in terms of network behavior. The principle of network analysis is what is at stake in the next subsections.

4.1 Term frequency as potential

The idea of a normalized term frequency can be expressed as a property of a network node. The example in Figure 1 shows five documents, of which two documents contain the term a, in both cases multiple times. In other words, the term frequency of a for documents I is 3 and 2 for document II. Looking at most ranking functions, this yields higher scores for document I and II, because the term frequency is often seen as a strong relevance indicator.

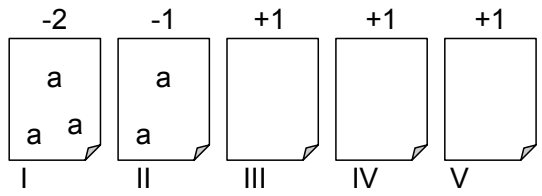

Figure 1. It is assumed that documents *I* and *II* are more relevant because of their higher term frequency w.r.t. documents *III* and up.

Continuing this thought, we might think of term frequency as the document's 'potential' to become the most relevant. If we consider the terms to be 'charges[1]', then document *I* and *II* tend to become more negative than the others. If the charges were evenly distributed, the potential of each of the documents had been zero (potential is a relative value). Now that the charge distribution is concentrated on the first two documents, the potential of all of the individual documents changes. Note that this approach shares properties with the normalized term frequency from section 3. The potential of an object is both influenced by the number of charges and the area of the object. The bigger the area, the lower the concentration of the charges, which decreases its potential. In this, the analogy with the document length can easily be found.

4.2 DF as a negative influence on potential

The inverse document frequency tells something about the importance of the term in question. The more common a term, the less weight should be addressed. Also, when a term does not occur in a document, the inverse document frequency is not relevant. In other words, the document frequency inversely contributes to the potential of the document. This can be considered a kind of 'resistance' that the term's score experiences. An absent term in a document corresponds with an infinite resistance for that term.

4.3 Putting together potential and resistance

The proposed measure of the relevance of a document based on potential and resistance is the basis of a well known rule in circuit analysis: Ohm's law. The rule is described by

$$I = V/R$$

By using this formula, it is possible to calculate the expected amount of current I that would flow from one network node (the document) to another (the query node). Depending on the potential of the nodes and the resistance of the vertices between them, the current varies.

[1]Since electrons are negative charges, so do we consider the terms

4.4 Network analysis

A flowing current between nodes caused by their difference of potential tells us something about the network. Using the same approach as Turtle and Croft, a query can be made a part of the dynamics of such network. In fig. 2 a single termed query is introduced as a central node between the other document nodes. The amount of current flowing from a document to the center can be seen as its ranking score.

Figure 2 shows such a network for a query containing the term a. Like in the example of fig. 1, documents *III*, *IV* and *V* do not contain the term in question. This is modeled by assigning an infinite resistance to resistors R_3, R_4, and R_5. The other resistors are given a resistance of 2 each[2]. That is, the number of documents in which a occurs. So in case term a could have occurred in 3 documents, each corresponding resistor would have been 3. The resistor at the bottom acts as a normalizer to the term. By adjusting its resistance, the net current is influenced.

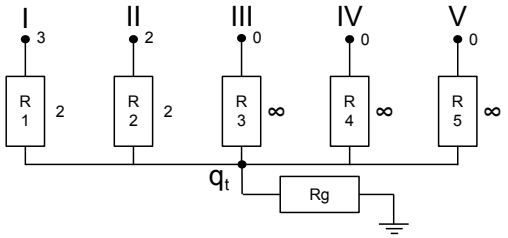

Figure 2. The term *t* of query *q* is part of the network (common node in the middle) and connected to documents *I* and *II*. The resistors R_1 and R_2 take the terms' document frequency into account.

For those who did not notice the circuit layout very well in fig. 2 (or for those who do not have an electronics background): the picture shows a typical voltage summer [Nilsson, 2001].

4.5 Multiple terms

The example in the previous section is based on a query containing a single term. Ideally, a user might want to issue multiple termed queries. Multiple query terms can be introduced as additional nodes to the network. The final score of a document is based on the superposition of all query terms. Each term contributes to the total score if it has a posting of the document in question. Every query term has its own link to a document by its own 'resistor'. This results in a current flowing to each individual query term node. Of course, the total score of the document-query combination must be a single value. A simple approach would be to sum the magnitudes of the term resistors:

$$R_+ = \sum_{i=1}^{N} s_i$$

[2]Dimensions are omitted since they are not of main concern

The final score of the document is then expressed by the current over one single resistor described by $I = V_d/R_+$. This is similar to how the score of the TF×IDF model is usually handled: the sum of the individual query term scores determines the final document score. This is however not a realistic approach to real network behavior. A central node representing the query must be used to connect each individual query term node. fig. 3 shows a circuit with this central node Q introduced.

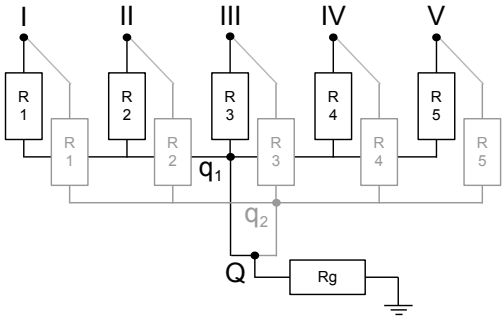

Figure 3. Query Q consists of multiple terms (q_1 and q_2) which are individually connected to document nodes.

This central node represents the query itself and is therefore connected to each query term node q_1 and q_2. Note that there are no resistors shown between the query node and each term node, but those can easily be introduced: by differentiation of these 'term resistors', the weight of the individual terms could be altered. Calculation of the net current flowing from the document nodes to the central query node requires more advanced network analysis methods. A well known heuristic is the parallel resistor theorem, which describes how to replace two or more parallel resistors for one equivalent resistor. In the situation in fig. 3, this means that the parallel resistors from Q to a document node d are substituted by an equivalent resistor, omitting the influence of other nodes. Substitution of two or more parallel resistors can be done by taking the inverse of the sum of the inverted values:

$$R_{||} = \left(\sum_{i=1}^{N} s_i^{-1} \right)^{-1}$$

This formula gives a more realistic value for the actual resistance acting between the central query node and the document. Figure 4 shows an example of resistor substitution for query Q consisting of two terms q_1 and q_2 for one document *(I)*.

5 Simulation

In section 4, we studied TF×IDF by trying to apply the laws of an electrical circuit to the model. Keeping a strict eye on the math of TF×IDF, there are clearly differences in the way a circuit behaves, although the physical view on the matter felt quite natural. The goal of this section is to discover the differences in precision of

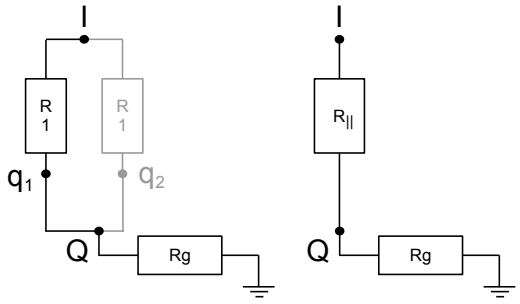

Figure 4. Example of resistor substitution of two resistors (left) for one equivalent resistor $R_{||}$ (right).

three models: TF×IDF, Ohm using summed scores, and Ohm using scores based on the equivalent resistance.

5.1 Test Setup

The evaluation was done using the Terrier IR Platform [Ounis et al., 2006] with indexed MEDLINE citations which were used during Text REtrieval Conference (TREC[3]) in 2005. Terrier has a built-in Java port of the TREC Evaluation program, which allows weighting models to be evaluated in terms of precision and recall, based on predefined topics and relevance feedback information. More on this topic follows in section 5.2. Besides an evaluation tool, Terrier also ships with various weighting models including a TF×IDF implementation. This model is out of the box already optimized with pre-set constants, differing from the model described in section 3. This model is defined as:

$$TF \times IDF = kf \frac{k_1 \cdot tf}{tf + k_1(1 - b + b \cdot dl/dl_{avg})} \log \frac{|D|}{df + 1}$$

Besides Terrier's predefined TF×IDF model used for evaluation, the Ohm model was implemented conform section 4 and added to Terrier's weighting models. This model is defined as:

$$Ohm = \frac{tf}{dl} df^{-1}$$

This formula can be derived by substitution into Ohm's formula given in section 4.3. The score is dictated by the current. The potential equals the term freqency ('charges') over the document length ('area') and the document frequency acts inversely: like a resistor. An explanation of symbols uesed can be found in table 1.

The Ohm model was evaluated using the two manners described in section 4.5. One is by summing the scores of the individual terms (denoted as 'Ohm single termed') and the second approach entails

[3] http://trec.nist.gov/

kf	KeyFrequency: the term frequency in the query
tf	The term frequency of the term in the document
$\|D\|$	The number of documents in the corpus
df	The document frequency of the term
dl	The document's length
k_1	Terrier constant: 1.2
b	Terrier constant: 0.75

Table 1. Symbols used in implemented WeightingModels

the parallel resistor approach to better express a realistic network behavior for multiple terms ('Ohm multiple termed').

5.2 Evaluation

A short look at the ranking models already reveals clear differences between both. First of all, the TF×IDF model is more complex in a sense that it has more terms, which include constants. These constants, k_1 and b, are used to fine-tune the model. Secondly, the TF×IDF model is proportional to the key frequency: the term frequency of the query. The Ohm model on the other hand, is neither influenced by properties by other terms in the query, nor by predefined constants.

It is to be expected that the TF×IDF will address the best precision, since it is already studied and evaluated during multiple TREC conferences. In order to quantify the differences between the models, a TREC evaluation was done for both models.

Figure 5 offers visual insight of the precision at x for both models by running an evaluation for all queries in one test run. Please note the log scale at the horizontal axis.

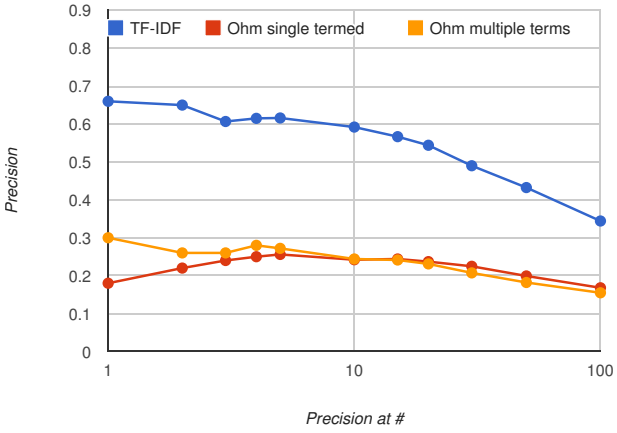

Figure 5. Evaluation of the models TF-IDF and Ohm (using scoring functions for both individual terms, and multiple terms combined).

6 Conclusions

Results show that the simplified Ohm network model cannot compete with the thoroughly optimized TF×IDF model. Although both models share similarities in sense of proportionality of term frequency and inverse document frequency, the additional terms of the TF×IDF model certainly cause a significant difference in precision. Despite that Ohm achieves a lower precision, the model might still be useful. The function result is easier to cache since it does not depend on the number of documents and the average document length. When a document is added to the corpus and index, only the terms involved need to be recalculated. In case of TF×IDF, the whole cache must be invalidated, which might not be an option for high-traffic websites with regularly changing content for example. Nonetheless this does not mean that the Ohm model can be considered a constant ranking model (see section 2.1). When calculating the score based on the parallel resistor theorem, the final score depends heavily on the combination of terms in the query. Throughout the results, the use of this scoring function for multiple terms combined does not achieve better results. Only the precision at the top 10 is slightly better compared to the other Ohm model which sums the individual term scores.

6.1 Differences and inconsistencies

The method used to calculate the net current flowing to the central query node is not as accurate as it could be. Due to limitations of Terriers design, it was significantly easier to adopt a heuristic approach, rather than using additional network simulation software. More realistic results require the use of Kirchhoff's laws (i.e. Kirchhoff's Current Law and Voltage Law). This requires specialized software and is certainly not trivial. In the field of electrical circuit analysis using software, simulation is more often used rather than exact numerical analysis.

6.2 Future work

This research only considers passive aspects network analysis, solely based on term frequency and document frequency as stated in section 4.4. No attempts were made to experiment with other properties of the voltage summer, for example by exchanging components or by introducing dynamic behavior. As made clear in section 6.1, looking at a ranking model by means of an electric network is more a thought experiment than a way to explain the 'physics' of information retrieval, as it can be done with real electric networks. Sequacious research should investigate the relevance and existence of more suitable network components to explain the behavior of information retrieval models. As depicted in the previous subsection, a network's behavior could for example be simulated instead of being calculated. The challenge would be to adjust existing software in this field to automatically generate large numbers of networks based on a collection of documents.

7 Acknowledgments

My special thanks go to Djoerd Hiemstra and Fokko Jan Dijksterhuis. Djoerd tutored an essential course in information retrieval and showed great patience during the long time it took me to finish my

research. Fokko Jan dared to delve into matters outside his usual field of research, which is in essence a very special core value of the Advanced Technology bachelor.

References

Norbert Fuhr. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Trans. Inf. Syst.*, 7(3): 183204, July 1989. ISSN 1046-8188. doi: 10.1145/65943.65944. URL http://doi.acm.org/10.1145/65943.65944 .

Djoerd Hiemstra. A probabilistic justification for using tfidf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131139, 2000. URL http://www.springerlink.com/index/965FRX0AHE63Q0Y0.pdf .

Djoerd Hiemstra. *Using language models for information retrieval*. Taaluitgeverij Neslia Paniculata, 2001. URL http://doc.utwente.nl/36473/1/t000001d.pdf .

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008. URL http://www.muict.polppolservice.com/Year3_2/IR/SEC3/L8[4slides].pdf .

James Nilsson. *Introductory circuits for electrical and computer engineering*. Prentice Hall Ptr, London, 2001. ISBN 9780130763686.

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of the OSIR Workshop*, page 1825, 2006. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.8824&rep=rep1&type=pdf#page=18 .

Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, October 2004. ISSN 0022-0418. doi: 10.1108/00220410410560582. URL http://www.emeraldinsight.com/journals.htm?articleid=864256&show=abstract .

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613620, November 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL http://doi.acm.org/10.1145/361219.361220 .

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988. ISSN 0306-4573. doi: 10.1016/0306-4573(88)90021-0. URL http://www.sciencedirect.com/science/article/pii/0306457388900210 .

Gerard Salton, Edward A. Fox, and Harry Wu. Extended boolean information retrieval. *Commun. ACM*, 26(11):10221036, November 1983. ISSN 0001-0782. doi: 10.1145/182.358466. URL http://doi.acm.org/10.1145/182.358466 .

Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):3543, 2001. URL http://ilps.science.uva.nl/Teaching/0405/AR/part2/ir_overview.pdf .

H. Turtle and W. B. Croft. Inference networks for document retrieval. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '90, page 124, New York, NY, USA, 1990. ACM. ISBN 0-89791-408-2. doi: 10.1145/96749.98006. URL http://doi.acm.org/10.1145/96749.98006 .

Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9(3): 187222, July 1991. ISSN 1046-8188. doi: 10.1145/125187.125188. URL http://doi.acm.org/10.1145/125187.125188 .