Why Language Models and Inverse Document Frequency for Information Retrieval?[☆]

Catarina Moreira, Andreas Wichert

Instituto Superior Técnico, INESC-ID
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal

Abstract

The issue of term weighting has been traditionally addressed in a heuristic way through TF.IDF. TF.IDF is a term weighting measure which has been developed as a heuristic. This measure can be seen as an information theoretical approach that adds all the information contained in a document set.

Statistical language models have been developed as a new form of automatically incorporating term frequencies and document length normalizations in a probabilistic form. This means that language models contain the TF.IDF measure in their framework and therefore they have nearly the same information content as the heuristic. This relation has been known in the information retrieval community, but researchers have been ignoring it. Many retrieval systems are built using complex probabilistic language models, where the simple TF.IDF heuristic could lead to similar performances.

In this paper, we review why statistical language models hold the same information content as TF.IDF. Although, these two approaches have different backgrounds, we review how related they are by transforming language models into TF.IDF through mathematical relations. In order to investigate the foundations of these two models, we examine the information theoretical framework through entropy formulas and mathematically derive TF.IDF. To derive language models, we also examine probabilistic functions and Naïve Bayes formulas. All these theoretical results were empirically tested on a dataset of academic publications from the Computer Science Domain. Results demonstrated that TF.IDF has approximately the same information content as statistical language models. This leaves the question of why using such complex probabilistic models, if similar performances could be achieved with the simple TF.IDF weighting measure.

Keywords: Mutual Information Gain, Entropy, Weighting Measures, Statistical Language Models, TF.IDF

1. Introduction

Most of the research work performed under the Information Retrieval domain is mainly based in the construction of retrieval models. Throughout the years, many models have been proposed to create systems which are accurate and reliable. The proposed systems range from the well-known vector space model [11, 21, 20, 5] to more probabilistic frameworks based on discriminative probabilistic models [7, 13, 20, 18] and language models [17, 3, 14, 8]. However, it has been verified in the literature that a retrieval model, by its own, is not capable to achieve a huge performance and consequently the usage of heuristics associated to documents and to individual terms was necessary.

The effectiveness of an information retrieval system is typically measured through the precision and recall metrics. Precision is the ratio of the number of relevant retrieved documents to the total number number of retrieved items. Recall, on the other hand, is the fraction of the number of relevant retrieved documents to the total of relevant documents in the collection [21].

[☆] This work was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013 and supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project PTDC/EIA-CCO/119722/2010

Email addresses: `catarina.p.moreira@ist.utl.pt` (Catarina Moreira), `andreas.wichert@ist.utl.pt` (Andreas Wichert)

Preprint submitted to Information Processing & Management *March 18, 2014*

In order to achieve proper performances over precision and recall, an information retrieval system must be able to return all documents which are likely to be relevant to the user and also be able to reject all documents which are not interesting for the user.

Following the work of [21], in order to enhance the recall metric, experiments have shown that terms that are frequently mentioned in individual documents need to be taken into account. This means that the term frequency factor (TF) has a substantial importance in a term-weighting system [20].

Term frequency alone is not enough to achieve plausible performances in retrieval systems. There are situations where the query terms are spread in the entire document collection, making the system retrieve all these documents and consequently affecting the precision of the results. This means that in order to fill the precision gap, a new factor must be introduced. That factor is the inverse document frequency (IDF). IDF is an heuristic which enables the discrimination of terms. Words that appear often in a collection of documents do not provide much information as words which occur occasionally. IDF is given by Equation 1 and is given by the logarithm of the inverse proportion of a word over the entire document corpus. In Equation 1, $|D|$ is the total number of documents in the collection and $|D_q|$ is the number of documents which contain the query term q.

$$IDF(q) = \log \frac{N}{N_q} \tag{1}$$

The combination of the term frequency measure and the inverse document frequency forms the well known $TF.IDF$, which is given by Equation 2. In this equation, $freq(q)$ is the number of times that the term q occurs in the document collection, N the total number of documents and N_q the number of documents that contain the terms q in their contents.

$$TF.IDF(q) = freq(q) \times \log \frac{N}{N_q} \tag{2}$$

Since TF.IDF has been developed as an heuristic, many researchers tried to find theoretical explanations of why this measure performs so well [19]. In the work of [2], TF.IDF can be seen as an Information Theoretical approach that adds all the information contained in a document set. Thus, it can be interpreted as the total quantity of information needed in order to compute the mutual information between documents and query topics. This means that TF.IDF can be thought as the reduction of the uncertainty about a random variable, representing the document collection, given the knowledge of another random variable, which represents the information need.

Along with TF.IDF, researchers in the Information Retrieval community also developed more complex retrieval systems based on discriminative probabilistic models or on statistical language models.

Statistical language models for information retrieval had also their foundations in Information Theoretic frameworks, through the work of Claude Shannon [23]. Shannon used n-grams combined with his entropy formula in order to investigate the information content of the English language. In Language Models, a document is represented as the probability of an ordered distribution of the vocabulary terms over the document. By assuming that each term is independent, then these approaches are simply based on the multiplication of the probability of each term being present in a document, and has its motivation in the Naïve Bayes formula. When applying the logarithm function, one can transform statistical language models into TF.IDF measures. This transformation indicates that the models are equivalent.

In this paper, we revise the transformations and mathematical relations between Language Models and TF.IDF. We analyze the information theoretic frameworks and Shannon's entropy formula and derive TF.IDF. We also examine probabilistic functions and Naïve Bayes formulas in order to derive the language models used in Information Retrieval. Although, these two approaches have different backgrounds, we show how related they are by transforming language models into TF.IDF. We also determine these theoretical results by empirical experiments over a dataset of academic publications from the Computer Science domain.

Many previous works have demonstrated elegant frameworks from which the TF.IDF heuristic could be derived so that its successful performance could be explained [16, 10, 19, 20]. The main goal of this paper is to demonstrate that statistical language models and TF.IDF can have the same information content. We therefore question why these measures are still being used together. Many researchers build retrieval systems based on complex language models.

If they have the same information content as the simple TF.IDF weighting measure, then why turning a system more complex by using them?

The rest of this paper is organized as follows: Section 2 presents the main concepts address in this work and that are crucial to understand this paper. Section 4 shows the results obtained in a simple empirical experiment where we compare TF.IDF against statistical language models and Shannon's Mutual Information Gain formula. In Section 4.1 we explain the similar results between TF.IDF and Mutual Information Gain by mathematical deriving TF.IDF from this information theoretic framework. In Section 4.2 we reveal the identical results between TF.IDF and Language Models and again we show that TF.IDF can be derived from this generative probabilistic model. Finally, Section 5 presents the main conclusions of this work.

2. Fundamental Concepts

This section presents all the concepts that will be used throughout this work and are crucial for understanding the mathematical derivations that will be performed in later sections. Since TF.IDF can be obtained from an information theoretic framework, we start this section by introducing the concepts of Entropy. Then, we provide a brief definition of Statistical Language Models for information retrieval and how we can derive the general formula through universal notions of probability theory.

2.1. Information Theory

Information theory is a field which is mainly based on probability theory and statistics. It also deals with the quantification of information. Most of the works concerned with information theory are based in the entropy formulations developed by Claude Shannon [22]. In this section, we present the main concepts related with entropy which need to be taken into account for further references in this work.

Shannon's Entropy: The entropy H of a discrete random variable X is a measure of the amount of uncertainty associated with the value of X [4]. Let X be discrete random variable with alphabet χ and probability mass function $P(x) = Pr[X = x], x \in \chi$. The entropy $H(x)$ of a discrete random variable X is defined by Equation 3.

$$H(X) = -\sum_{x \in \chi} P(x) \log P(x) \tag{3}$$

Just like Shannon's entropy, IDF can also be seen in a probabilistic perspective in the following way. The probability of random document d containing a query term q can be approximately given by the following formula [10].

$$P(q) = P(\texttt{q occuring in d}) \approx \frac{N_q}{N}$$

So IDF can be redefined in terms of probability theory in the following way (note that $\log \frac{1}{x} = -\log x$).

$$IDF(q) = -\log P(q)$$

When considering more than one term, IDF can be given by simply summing the individual IDF scores of each query term [19].

$$\begin{aligned} idf(t_1 \cap t_2) &= -logP(t_1 and t_2) \\ &= -\log P(t_1)P(t_2) \\ &= -(\log P(t_1) + \log P(t_2)) \\ &= idf(t_1) + idf(t_2) \end{aligned}$$

Finally, one can already notice some relation between the IDF measure and Shannon's entropy, since they share the same logarithmic part in their formulas.

$$H(q) = P(q)IDF(q)$$

Conditional Entropy: We define the conditional entropy of a random variable given another as the expected value of the entropies of the conditional distribution, averaged over the conditioning random variable. If $(X, Y) \sim P(x, y)$, the conditional entropy $H(Y|X)$ is defined by Equations 4-6, [4].

$$H(Y|X) = \sum_{x \in \mathcal{X}} P(x) H(Y|X = x) \tag{4}$$

$$= -\sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} P(y|x) \log P(y|x) \tag{5}$$

$$= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log P(y|x) \tag{6}$$

Mutual Information Gain: Is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other [4]. Consider two random variables X and Y with a joint probability mass function $P(x, y)$ and marginal probability mass function $P(x)$ and $P(y)$. The mutual information $I(X; Y)$ is the relative entropy between the joint distribution and the product distribution $P(x)P(y)$.

$$I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) log \frac{P(x, y)}{P(x)P(y)} \tag{7}$$

$$= \sum_{x,y} P(x, y) log \frac{P(x|y)}{P(x)} \tag{8}$$

$$= -\sum_{x,y} P(x, y) \log P(x) + \sum_{x,y} P(x, y) \log P(x|y) \tag{9}$$

$$= -\sum_{x} P(x) log P(x) - \left(-\sum_{x,y} P(x, y) log P(x|y) \right) \tag{10}$$

$$= H(X) - H(X|Y) \tag{11}$$

3. Language Models for Information Retrieval

In Statistical Language Models, a document is a good match for a query if a probabilistic generative model for the documents is capable of generating the query, which happens when the document contains the terms of the query more often. Statistical Language Models build a model θ_d from each document d and thereafter rank the documents based on the probability of the document model having generated the query, i.e. $P(q|\theta_d)$ [12].

Let $D = d_1, d_2, ..., d_n$ be a set of documents and $T = t_1, t_2, ..., t_m$ be a set of distinct terms contained in the documents. Given the query terms, we are interested to know the probability which is assigned to the documents D. If we assume that the query terms are conditionally independent given the documents, then we can represent this probability through Naïve Bayes formula.

$$P(D|t_1, t_2, ..., t_m) = \frac{P(t_1, t_2, ..., t_m|D).P(D)}{P(t_1, t_2, ..., t_m)}$$

Since the denominator $P(t_1, t_2, ..., t_m)$ does not depend on the documents, then it has a constant value which will not interfere in the ranking process. So, we can ignore it. The prior probability $P(D)$ can also be ignored if and only if we assume that all documents have the same probability of being relevant when no query topic is provided. Assuming that $P(D) = 1/n$, so we are supposing that the documents are all equally likely relevant and therefore we can ignore the $P(D)$ term in the formula [8].

In order to construct the language model of the documents, we still need to determine the probability of the query terms, given the documents, $P(t_1, t_2, ..., t_m|D)$. To build such model, each document of D is represented as the probability of an ordered distribution of the vocabulary terms over the document. This is represented through m random variables. If we model the terms of query and of a document as compound events (i.e. an event which consists of two or more events), then we can assume that these events are independent and consequently we obtain the following formula:

$$P(t_1, t_2, ..., t_m|D) = \prod_{i=1}^{m} P(t_i|D) \tag{12}$$

Note that the above model cannot be seen as a real probability value, since the values are not normalized, therefore we shall address to this values rather as a score than a probability. The above model is usually referred as the query likelihood model scores and ranks documents based on the probability assigned to the query using their individual language models. Since we are multiplying the terms' probabilities, one can notice that the longer the query, the lower will be the document scores.

The above formula has the disadvantage of returning a zero score if one query topic is not present in the document set. Smoothing techniques based on linear interpolations can solve the problem by decreasing the probability of observed events and by increasing the probability of unseen outcomes. So,Equation 12 can be rewritten as:

$$score(Q|D) = P(t_1, t_2, ..., t_m|D) = \prod_{i=1}^{m} (\lambda_i P(t_i) + (1 - \lambda_i) P(t_i|D)) \tag{13}$$

In Equation 13, λ corresponds to a smoothing parameter which is usually set to 0.5 and the term Q corresponds to a query with m terms, $Q = \{t_1, t_2, \ldots, t_m\}$.

4. A Simple Experiment

In this section, we perform a simple experiment where we compare the retrieval performance of TF.IDF against standard statistical language models and Shannon's mutual information gain. In order to validate the proposed experiment, we required a sufficiently large repository of textual information. The dataset chosen was a public available database with academic publications from the Computer Science domain, the DBLP database[1]. This dataset is a very rich and contains 1 632 440 publication's from which 653 511 contain also the publication's abstracts. For our experiments we collected these half million documents with the abstract information, in order to perform the retrieval process.

The preference for this dataset was simply because the authors already have a good knowledge of its structure. The DBLP dataset has also been widely used in the information retrieval community for citation analysis [24] and to find academic experts [25, 15].

We manually made 35 query topics based on computer science topics which can be found in the publications covered in the DBLP dataset. Some of these queries have already been used in other tasks of information retrieval [6, 25]. Table 1 shows the queries used for our experiment.

Figure 1 shows the scores obtained using the TF.IDF measure (Equation 2), statistical language models (Equation 13) and Shannon's Mutual Information Gain formula (Equation 7).

In order to make the results of these three formulas equivalent, we needed to convert the statistical language models from Equation 13 into a sum of logarithmic functions. Since in the DBLP dataset we are dealing with a reasonably large amount of terms, multiplying the probabilities of each term will lead to very low results. This would avoid us to fairly compare unnormalized measures such as TF.IDF against the statistical language models. Equation 14 shows the equation used in our experiment for these models. This formula will avoid the low probability results achieved by the

[1] http://www.arnetminer.org/citation

Business Intelligence	Ontologies	Bayesian Networks	Natural Language
Signal Processing	Support Vector Machines	Dimension Reduction	Question Answer
Information Retrieval	Quantum Computation	Sensor Fusion	Neural Networks
Spatial Cognition	Game Theory	Indexing Structures	Mobile Networks
Computer Vision	Boosting	Human Computer Interaction	Computer Architecture
Decision Making	Cryptography	Information Theory	Neuroimaging
Artificial Intelligence	Machine Learning	Automata	Enterprise Architecture
Associative Memory	Expert Search	Computer Graphics	Distributed Parallel Systems
Intelligent Agents	Geographic IR	Information Extraction	

Table 1: Queries used in order to simulate the proposed experiments.

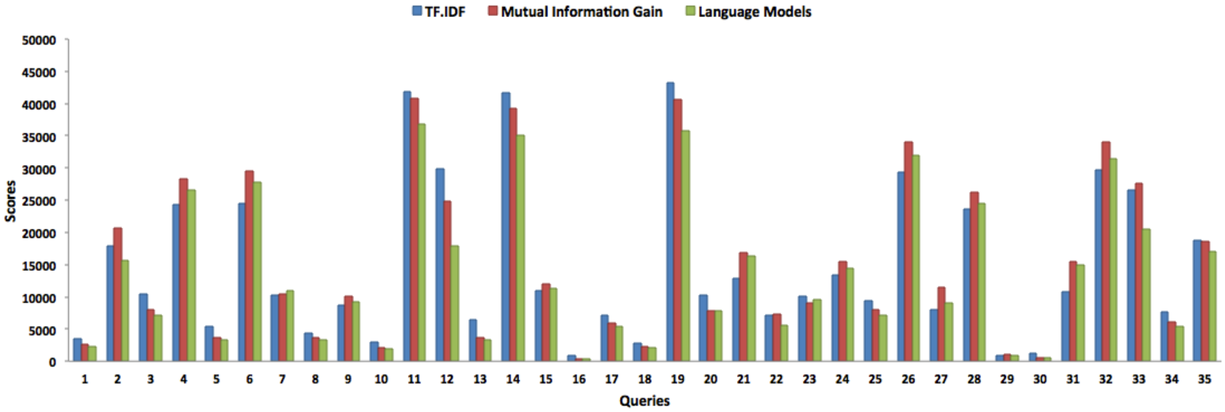

Figure 1: Results obtained using the queries specified in Table 1 and three different document weighting measures: TF.IDF, statistical language models and mutual information gain.

standard language models formula (Equation 13), making it possible to compare its results with the TF.IDF and the mutual information gain formulas. Equation 14 contains a parameter λ that needs to be manually tuned. We used the values that the general works of information retrieval use, that is, we set $\lambda = 0.5$.

$$score(Q|D) = P(t_1, t_2, ..., t_m|D) = \sum_{i=1}^{m} \log(\lambda_i P(t_i) + (1 - \lambda_i)P(t_i|D)) \qquad (14)$$

Following Figure 1, one can clearly see that these three formulas achieve approximately the same performance. We also performed a paired t test in order to determine if the three methods are statistically significant. Results showed that the significance tests performed did not accuse any differences between the three methods (that is, they are not statistically significant). This means that the three approaches achieve similar performances. In the following sections, we explain mathematically why these results are so similar. In fact, we will show that TF.IDF can be derived from an information theoretic framework such as the mutual information gain formula [2]. We will also demonstrate that, although standard language models do not have an explicit use of TF or IDF, it turns out that the TF.IDF weighting measure can be derived from such probabilistic models [8].

4.1. Relation Between TF.IDF and Mutual Information Gain

In this section, we explain by revising the work of [2] the similar results obtained between TF.IDF and the mutual information gain formula.

Let $D = d_1, d_2, ..., d_N$ be a set of documents and $Q = t_1, t_2, ..., t_M$ be a set of distinct terms contained in the documents. We are interested in finding the total amount of information that a document contains over some query terms. more specifically, we are interested in the reduction of the amount of uncertainty of a document due to the knowledge that it contains the query terms. This is given by the mutual information gain formula.

$$I(D;T) = H(D) - H(D|T)$$

The total amount of information that the random variable D, which represents the entire document set of a collection, contains is given by its self entropy, that is:

$$H(D) = -\sum_{d_j \in D} P(d_j) \log P(d_j)$$

Assuming that all documents are equally likely to be retrieved, then we can state that the probability of a document is given by $1/N$ and therefore the above statement becomes:

$$H(D) = -\sum_{d_j \in D} P(d_j)\ logP(d_j) = -N\frac{1}{N} \log \frac{1}{N} = -\log \frac{1}{N}$$

We can multiply the above statement by $\sum_{t_i \in T} P(t_i)$. Since the probability of the query terms is always the same for every document, then adding this information to the formula will not affect the entropy of the random variable D, because we are multiplying it by a constant. This will be useful to facilitate some calculations later.

$$H(D) = -\sum_{t_i \in T} P(t_i) \log \frac{1}{N} \tag{15}$$

In order to compute the mutual information gain, the conditional entropy of a document given the query terms, $H(D|T)$ has to be computed. This formula means that only the subset of documents containing the query terms t_i are considered.

$$H(D|T) = -\sum_{t_i \in T} P(t_i) \sum_{d_j \in D} P(d_j|t_i) \log P(d_j|t_i) \tag{16}$$

In Equation 16,assuming that the N_t documents are equally likely, the amount of information calculated for each document in the subset is $-\log(\frac{1}{N_t})$. This gives:

$$\begin{gathered} H(D|T) = -\sum_{t_i \in T} P(t_i) \sum_{d_j \in D} P(d_j|t_i) \log P(d_j|t_i) = -\sum_{t_i \in T} P(t_i) N_t \frac{1}{N_t} \log \frac{1}{N_t} \\ H(D|T) = -\sum_{t_i \in T} P(t_i) \log \frac{1}{N_t} \end{gathered} \tag{17}$$

From this point, we already have all the information required to compute the mutual expectation information gain. We just need to make the difference between Equation 15 with Equation 17.

$$I(D;T) = H(D) - H(D|T)$$

$$I(D;T) = -\sum_{t_i \in T} P(t_i) \log \frac{1}{N} + \sum_{t_i \in T} P(t_i) \log \frac{1}{N_t}$$

$$I(D;T) = \sum_{t_i \in T} P(t_i)(-\log \frac{1}{N} + \log \frac{1}{N_t})$$

$$I(D;T) = \sum_{t_i \in T} P(t_i)(\log N - \log N_t) = \sum_{t_i \in T} P(t_i) \log \frac{N}{N_t}$$

In the above statement, $P(t_i)$ is the probability of the query term t_i. $P(t_i)$ is given by the frequency of the query term t_i in the whole document set divided by the total number of terms in the entire document set.

$$P(t_i) = \frac{\sum_{d_j \in D} freq(t_i, d_j)}{\sum_{d_j \in D} \#terms(d_j)}$$

Substituting this statement in the previous formula, we obtain:

$$I(D;T) = \sum_{t_i \in T} \sum_{d_j \in D} \frac{freq(t_i, d_j)}{\#terms(d_j)} \log \frac{N}{N_t} \tag{18}$$

In this point, Equation 18 contains some terms that resemble the TF.IDF formula [2]:

- $freq(t_i, d_j)$ is the frequency of the term t_i in document d_j, also known as TF of term t_i;
- $\log \frac{N}{N_t}$ is the inverse document frequency, where N corresponds to the total number of documents in the document set and N_t is the number of documents which contains the query topics.
- $\#terms(d_j)$ is a normalization factor which is not used in the traditional TF.IDF formula.

Given this, one can conclude that the TF.IDF values do not represent a probability value, but can rather be interpreted as the quantity needed for the calculation of the expected mutual information gain. One can also note that when deriving the TF.IDF formula, it was assumed an equal probability to all documents containing the query terms. Under the information theory framework this assumption has the consequence of maximizing the entropy values, making TF.IDF one of the most important weighting measures of the Information Retrieval literature [1].

4.2. Relation Between TF.IDF and Statistical Language Models

Statistical Language Models can be interpreted as TF.IDF weighting algorithm with document normalization. In this section, we will review the work of [8] so that we can show TF.IDF can be derived from statistical language models. In order to demonstrate this, we will start by the definition a Language Model which was already presented in Equation 13.

$$score(Q|D) = \prod_{q_i \in Q} (\lambda_i P(q_i) + (1 - \lambda_i) P(q_i|D))$$

If we multiply the above formula by 1, will not affect the ranking of the documents.

$$score(Q|D) = \prod_{q_i \in Q} (\lambda_i P(q_i) + (1 - \lambda_i) P(q_i|D)) \frac{\lambda_i P(q_i)}{\lambda_i P(q_i)} \tag{19}$$

$$score(Q|D) = \prod_{q_i \in Q} \left(\frac{\lambda_i P(q_i)}{\lambda_i P(q_i)} + \frac{(1 - \lambda_i) P(q_i|D)}{\lambda_i P(q_i)} \right) \lambda_i P(q_i) \tag{20}$$

$$score(Q|D) = \prod_{q_i \in Q} \left(1 + \frac{\lambda_i P(q_i)}{(1 - \lambda_i) P(q_i|D)} \right) \lambda_i P(q_i) \tag{21}$$

Since $(1 - \lambda_i) P(q_i|D)$ is a constant, it does not affect the ranking of the documents, and therefore it can be ignored [8] .

$$score(Q|D) = \prod_{q_i \in Q} \left(1 + \frac{\lambda_i P(q_i)}{(1 - \lambda_i) P(q_i|D)} \right) \tag{22}$$

To approximate our formula to the TF-IDF weighting algorithm, it would be useful getting rid of the products. This can be done by using summations and logarithms in the following way:

$$score(Q|D) = \sum_{q_i \in Q} \log \left(1 + \frac{\lambda_i P(q_i)}{(1 - \lambda_i) P(q_i|D)} \right) \tag{23}$$

We know that the probability of of the query terms is equal to the document frequency of the query term q_i, divided by the whole set of terms in the document set, that is $P(q_i) = \frac{df(q_i)}{\sum_{q \in Q} df(q)}$. On the other hand, the probability of a query term given a document is $P(q_i|D = d_j) = \frac{tf(q_i,d_j)}{\sum_{q \in Q} tf(q,d_j)}$. So the above formula becomes:

$$score(Q|D) = \sum_{d_j \in D} \sum_{q_i \in Q} \log \left(1 + \frac{\lambda_i}{(1 - \lambda_i)} \frac{tf(q_i, d_j)}{df(q_i)} \frac{\sum_{q \in Q} df(t)}{\sum_{q \in Q} tf(q, d_j)} \right) \quad (24)$$

In Equation 24, one can notice that [8]:

- $\frac{tf(q_i,d_j)}{df(q_i)}$ can be seen as the TF.IDF weight of the query term q_i in the document d_j.
- $\frac{\lambda_i}{(1-\lambda_i)}$ is the odds of the probability of term importance given relevance.
- $\frac{1}{\sum_{q \in Q} tf(q,d_j)}$ is the inverse length of document d_j.
- $\sum_{q \in Q} df(q)$ is constant for any document d and term q. This value needs only to be computed once for the entire document collection.

5. Conclusion

In this paper, we showed some mathematical foundations for the TF.IDF weighting measure and for the statistical language models. TF.IDF can be derived from an information theoretical approach that adds all the information contained in a document set. Thus, it can be interpreted as the total quantity of information needed to reduce the uncertainty about a document random variable given the knowledge of a query random variable. The statistical language models, on the other hand, can be derived from probabilistic functions and Naïve Bayes formulas.

In this paper, we demonstrated that, although these two approaches have different backgrounds, they are related by transforming language models into TF.IDF through mathematical relations. We also validated these theoretical findings through empirical experiments on a database of academic publications from the Computer Science Domain. The results obtained were in accordance with our theoretical mathematical hypothesis, that is TF.IDF can have been derived from an information theoretic framework and statistical language models have a similar information content as the TF.IDF weighting measure.

After this demonstration, one might be thinking that if statistical language models can be derived into TF.IDF, then what are the advantages of using these models over the TF.IDF weighting measure? According to [9], statistical language models have the ability of representing TF.IDF through probability theory. This can become very helpful, since it enables the computation of the probability of randomly choosing the query terms, one at a time, from a document. Thus, one can model complex information retrieval queries in a simple and elegant manner. However, these models have disadvantages over the precision of the results. The probability of many events converges to very small values. Thus, language models should not be used in the product form if many terms are required to perform the calculations. In such situations, converting the multiplications by a sum of logarithms should be more advisable just like we did in our empirical experiments.

References

[1] A. Aizawa, The freature quantity: An information theoretic perspective of tf.idf-like measures, in: Proceedings of the 23rd Annual Internatinal ACM SIGIR Conference on Research and Development in Information Retrieval.

[2] A. Aizawa, An information-theoretic perspective of tf-idf measures, Information Processing Management 39 (2003) 45–65.

[3] A. Berger, J. Lafferty, Information retrieval as statistical translation, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[4] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley & Sons Inc., 2006.

[5] B. Croft, D. Harper, Using probabilistic models of document retrieval without relevance information, Journal of Documentation 35 (1979) 285–295.

[6] H. Deng, I. King, M.R. Lyu, Enhanced models for expertise retrieval using community-aware strategies, Journal of IEEE Transactions on Systems, Man, and Cybernetics 99 (2011) 1–14.

[7] N. Fuhr, Probabilistic models in information retrieval, Computer Journal 35 (1992) 243–255.

[8] D. Hiemstra, A linguistically motivated probabilistic model of information retrieval, in: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries.

[9] D. Hiemstra, Statistical language models for intelligent xml retrieval, in: Intelligent Search on XML Data: Applications, Languages, Models, Implementations, and Benchmarks, Springer, 2003.

[10] K.S. Jones, A statistical interpretation of term specificity and its applications in retrieval, Journal of Documentation 28 (1972) 11–21.

[11] H.P. Luhn, A statistical approach to the mechanized encoding and searching of literary information, IBM Journal of Research and Develpment 1 (1957) 309–317.

[12] C.D. Manning, Introduction to Information Retrieval, Cambridge University Press, 2008.

[13] M.E. Maron, J.L. Kuhns, On relevance, probabilistic indexing, and information retrieval., Journal of ACM 7 (1960) 216–244.

[14] D. Miller, T. Leek, R. Schwastz, A hidden markov model information retrieval system, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[15] C. Moreira, P. Calado, B. Martins, Learning to rank for expert search in digital libraries of academic publications, in: Progress in Artificial Intelligence, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, pp. 431–445.

[16] K. Papineni, Why inverse document frequency?, in: Proceedings of the North American Chapter of the Association for Computational Linguistics.

[17] J. Ponte, B. Croft, A language modelling approach to information retrieval, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[18] C. van Rijsbergen, Information Retrieval, Butterworths, 1979.

[19] S. Robertson, Understanding inverse document frequency: On theoretical arguments for idf, Journal of Documentation 60 (2004) 503–520.

[20] S.E. Robertson, K.S. Jones, Relevance weighting of search terms, Journal of the American Society for Information Science 27 (1976) 129–146.

[21] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Journal of Information Processing and Management 24 (1988) 513–523.

[22] C. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423.

[23] C. Shannon, Prediction and entropy of printed english, Bell System Technical Journal 30 (1951) 50–64.

[24] A. Sidiropoulos, Y. Manolopoulos, A citation-based system to assist prize awarding, Journal of the ACM Special Interest Group on Management of Data Record 34 (2005) 54–60.

[25] Z. Yang, J. Tang, B. Wang, J. Guo, J. Li, S. Chen, Expert2bole: From expert finding to bole search, in: Proceedings of the 15th ACM Conference on Knowledge Discovery and Data Mining.