

Document Length Normalization by Statistical Regression

Sylvain Lamprier, Tassadit Amghar, Bernard Levrat and Frederic Saubion
LERIA, University of Angers
2, Bd Lavoisier 49000 ANGERS, FRANCE
{lamprier, amghar, levrat, saubion}@info.univ-angers.fr

Abstract

The document-length normalization problem has been widely studied in the field of Information Retrieval. The Cosine Normalization [2], the Maximum tf Normalization [1] and the Byte Length Normalization [12] are the most commonly used normalization techniques. In [14], authors studied the retrieval probability of documents w.r.t. their size, using different similarity measures. They have shown that none of existing measures retrieve the documents of different lengths with the same probability. We first show here that the document and query sizes are indeed very influent on the similarity score expectation. Therefore, we propose to realize a statistical regression of the similarity scores distribution w.r.t. document and query sizes in order to normalize them. Experimental results appear to indicate that our approach, as well in the field of classical Information Retrieval as when applied to a document clustering process, allows to judge similarities really more fairly.

1 Introduction

An Information Retrieval system typically returns, as response to a user's query, a ranked list of documents [16]. In order to build this list, the probability, for each document, of being relevant for the request, has to be estimated. Several measures have then been proposed to compute the similarity between documents and queries [19]. Most of them rely on the vectorial model [2], on the probabilistic model [12] or on simple co-occurrence of terms. Term¹ weighting is an important aspect of these models since every terms have not the same importance in a document. Three main factors affect the weight of a term in a text: the term frequency factor (*tf*), the inverse document frequency factor (*idf*) and the document length normalization [13]. Whereas the *tf* factor corresponds to the number of occurrences of a term within

¹Terms are indexing units used to identify the topics of a text. In this paper, units used are meaningful words of the texts, after having removed the stop words and applied a stemming process [11].

the text, the *idf* factor renders its ability of discrimination. Since longer documents contain more terms, their probability to contain query's terms is higher. Same manner, the terms of these documents may more probably be repeated several times, query's terms are thus likely to have a better frequency in the text. Document length normalization aims to reduce the advantage of that long documents.

The document-length normalization problem has been widely studied in the litterature. The Cosine Normalization [2], the Maximum tf Normalization [1] and the Byte Length Normalization [12] are the most commonly used normalization techniques. In [14], authors have studied the probability of retrieval of each document w.r.t. its size, using the three similarity measures which have obtained the best results in TREC-3 [6]. They have shown that none of these measures retrieve documents of different lengths with a same probability. Assuming that longer documents have more chances to be relevant, they have proposed a new normalization technique, called the Pivoted Document Length Normalization. An investigation of the assumption realized led us to search for a new normalization function that relies on a statistical regression of the similarity scores distribution w.r.t. document and query sizes.

Document collections used for our experiments are described in Section 2. Section 3 presents some similarity measures and Section 4 investigates the Pivoted Document Length Normalization proposed in [14]. Then, Section 5 addresses the adaptation of the retrieval probability to the relevance probability and proposes a new evaluation measure of the effectiveness of retrieval systems that considers the length of retrieved documents. A study of similarity scores is realized in Section 6 and our normalization technique is proposed in Section 7. Finally, this normalization is applied to compute inter-text similarities in Section 8, by studying its impact in the field of document clustering.

2 Document Collections

Four document collections taken from the TREC corpus were used in the experiments. The ZIFF corpus in-

cludes computer science articles copyrighted by Ziff-Davis Publishing Company, FR whole issues of the Federal Register, and AP and WSJ full articles from the Associated Press and the Wall Street Journal respectively. Table 1 presents

Document collection	ZIFF	AP	WSJ	FR
Number of documents	75180	84510	98732	25960
Mean number of terms per document	297	217	204	927
Mean number of unique terms per document	139	144	128	244
Mean number of terms of the 1000 shortest documents	19	14	11	49
Mean number of terms of the 1000 longest documents	7915	512	1235	11304
Mean number of unique terms of the 1000 shortest documents	14	13	12	35
Mean number of unique terms of the 1000 longest documents	1503	323	623	1392
Mean relevant documents per query	54.06	40.44	54.26	19.32

Table 1. Statistics of the collections

the statistics of these corpuses. The number of unique terms is obtained by keeping a single occurrence of each term in each document. The same set of queries, the topics 1-50 of TREC, is used for each corpus. It should be noted that each corpus owns a great variety of document sizes. Whereas the FR corpus contains a lot of very long documents, WSJ and AP collections contain rather short ones. The ZIFF corpus presents the greatest heterogeneity.

3 Main Similarity Measures

In this section, we describe the three retrieval systems that have obtained the best results in TREC-3, i.e., Cornell’s Smart [13], Okapi [12] and Inquery [1]. These statistical systems rely on a *tfθidf* term weighting scheme. Smart is based on a classical vectorial model [2] where documents are encoded in vectors of weights w.r.t. the set of their meaningful terms and the similarity of a document with a query is computed by a cosine measure. Inquery relies on the inference network model and uses a probabilistic technique to determine the term weights. Okapi uses an approximation of the 2-Poisson model to evaluate the importance of a term in a document. In [14], authors established approximations of Okapi [12] and Inquery [1] in order to transpose them into a vectorial model, allowing then a unique similarity computation for the three measures:

$$Sim(D, Q) = \sum_{i=1}^T W_{D_i} \times W_{Q_i} \quad (1)$$

where T is the number of meaningful terms and W_{D_i} and W_{Q_i} the weights of the term i in the document and the query respectively. In the following, tf_i represents the term frequency of the i th term in the document/query text, N the total number of documents in the collection, n_i the number of documents containing the term i , max_{tf} the maximum term frequency within the document, and dl and $avdl$ the document length (in bytes) and its average respectively. According to the measure, W_{D_i} and W_{Q_i} correspond to:

Smart:

$$W_{D_i} = \frac{w_{D_i}}{\sqrt{\sum_{j=1}^T w_{D_j}^2}}, w_{D_i} = 1 + \log(tf_i)$$

$$W_{Q_i} = \frac{w_{Q_i}}{\sqrt{\sum_{j=1}^T w_{Q_j}^2}}, w_{Q_i} = (1 + \log(tf_i)) \times \log \frac{N}{n_i}$$

Okapi:

$$W_{D_i} = (tf_i \times \log(\frac{N-n_i+0.5}{n_i+0.5})) / (2 \times (0.25 + 0.75 \times \frac{dl}{avdl}) + tf_i)$$

$$W_{Q_i} = tf_i$$

Inquery:

$$W_{D_i} = 0.4 + \frac{0.6 \times \log(\frac{N}{n_i})}{\log(N)} \times (0.4 \times H + \frac{0.6 \times \log(tf_i+0.5)}{\log(max_{tf}+1.0)})$$

$$\text{with } H = 1.0 \text{ if } max_{tf} \leq 25, H = \frac{25}{max_{tf}} \text{ otherwise}$$

$$W_{Q_i} = tf_i$$

The effectiveness of the similarity measures is evaluated w.r.t. their ability to rank the relevant documents at the top of the list of documents retrieved. Two criteria, based on the classical precision measure *Prec*, which computes the ratio of relevant documents within the set of the n first documents retrieved, are used to assess this ability: the mean average precision, *MAP*, which realizes the average of the precisions computed after each relevant document of the list, and the precision at rank 100, *P@100*, which computes the precision after 100 documents retrieved. Table 2 gives

Classical Measures	Smart				Okapi				Inquery			
	Title		Narrative		Title		Narrative		Title		Narrative	
	MAP	P@100	MAP	P@100	MAP	P@100	MAP	P@100	MAP	P@100	MAP	P@100
ZIFF	0.176	0.123	0.315	0.209	0.187	0.140	0.309	0.227	0.137	0.115	0.111	0.127
AP	0.150	0.098	0.330	0.209	0.175	0.114	0.328	0.208	0.158	0.105	0.251	0.163
WSJ	0.151	0.128	0.383	0.275	0.215	0.158	0.405	0.278	0.187	0.153	0.260	0.193
FR	0.129	0.044	0.303	0.119	0.149	0.062	0.312	0.109	0.137	0.044	0.128	0.061

Table 2. Effectiveness of the measures

the average of the results obtained w.r.t. to these two criteria over each corpus with the set of queries mentioned in Section 2. In this table, Title refers to experiments realized by only using title (keywords) of queries (an average of 2.95 meaningful terms by query, 2.95 unique terms) and Narrative to results obtained with the full narrative description of queries (an average of 88.08 meaningful terms by query, 52.6 unique terms)². The results obtained show, as it is the case in TREC-3 [6], the dominance of Okapi when using the titles of the queries. Nevertheless, the effectiveness of Smart appears to really increase with longer queries. On the other hand, Inquery seems to be only suited to retrieve documents as response to short queries.

4 Adapting Retrieval Probability w.r.t. Relevance Probability

In [14], authors have proposed to sort documents in a collection w.r.t. their length and to divide them into bins in order to compare the probabilities of retrieval and relevance of documents w.r.t. to their size (their number of meaningful terms). In this way, the 75000 shortest documents of the ZIFF corpus have been divided into 84 bins of 1000

²Title and Narrative correspond to fields names of the TREC’s topics.

documents. Same manner as in [14], we have studied the distribution, over the created bins, of the first 1000 documents retrieved, with each similarity measure, for each of the 44 queries having relevant documents in the ZIFF corpus. The retrieval probability of documents of a given bin is computed by dividing the number of documents of that bin figuring in the lists of the first 1000 documents retrieved as response to each query by the total number of documents retrieved (44000 documents).

The probability of finding a relevant document of a given length is estimated by the number of documents in each bin, being relevant for each of the 44 queries, divided by the total number of relevant documents (2703 documents).

Figure 1 plots the distribution of retrieval and relevance probabilities w.r.t. to the bin lengths³. Curves of the graph on the left have been obtained by using Title queries and curves of the graph on the right by using the Narrative queries. These graphs first highlight that long documents

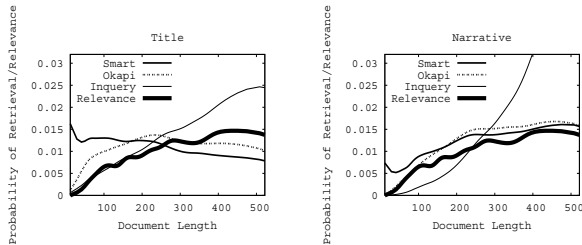


Figure 1. Retrieval/Relevance probabilities

are more frequently relevant than short ones. In [14], authors claimed that the better effectiveness of Okapi is induced by a better fit of the retrieval probability of documents of a given length with the probability of finding a relevant document of that length. With the Smart measure, the retrieval probability is greater than the relevance probability for short documents and lower for long ones when considering title queries. [14] thus proposed to modify this measure in order to lower the retrieval probability of short documents and to increase it for long ones. The basic idea is consequently to pivot and tilt the normalization factor of the document term weights used in the Smart system. In that way, a pivot point and a slope degree are chosen to compute the new document term weights:

$$W_{D_i} = \frac{w_{D_i}}{(1 - slope) \times pivot + slope \times \sqrt{\sum_{j=1}^T w_{D_j}^2}} \quad (2)$$

In [14], authors suggest to set the pivot equal to the average of the old normalization factor ($\sqrt{\sum_{j=1}^T w_{D_j}^2}$) computed over every documents of the corpus. They realized experiments to fix an optimal value for the slope. After training, this slope is set to 0.2. Nevertheless, in [4], experiments

³The length of a bin corresponds to the average length of its documents.

have shown that the size of the query influences the optimal slope. The graph on the right of Figure 1 confirms this observation since curves are different from those obtained when considering only the title of the queries. In [4], authors have therefore studied the evolution of the optimal slope w.r.t. the size of the query and determined an equation to estimate it: $slope = 0.0921 \times \log(QL) + 0.0658$, with QL being the number of unique terms of the query. Table 4

Pivoted Smart	Title		Narrative	
	MAP	P@100	MAP	P@100
ZIFF	0.159	0.166	0.301	0.227
AP	0.174	0.117	0.344	0.220
WSJ	0.223	0.165	0.408	0.278
FR	0.173	0.060	0.307	0.090

Table 3. Pivoted Smart results

presents the results obtained with the pivoted normalization applied to the Smart measure, using this estimated slope. According to these results, the pivoted normalization allows to widely improve the Smart results.

5 Why Are Longer Documents More Frequently Relevant?

In [14], authors claimed that longer documents are more useful for the user than short ones since they correspond to a greater number of queries. The fact that long documents are more often relevant than short ones appears to be valid (see Figure 1). It could be explained by the fact that longer documents are related to more topics. Nevertheless, it cannot render the utilisability of such documents by the user. Indeed, the skimming of the information in long documents is more difficult and implies to furnish a more important cognitive effort to find the interesting information. In TREC, the judgements of relevance realized are binary, a document is judged relevant or not and there is no available information about the relevance degree. Relevant judged documents are thus as likely to be interesting for the information needs of the user, independently of their size. For Jacobs [7], users prefer direct answers to queries rather than long texts containing relevant information. A short relevant document may thus be really more interesting for the user than a long one in which information has to be extracted. Moreover, with a normalization technique that favors long documents, such as the pivoted normalization [14], the non-relevant documents retrieved are likely to be longer too. Consequently, the identification of these documents to brush aside them is more difficult.

On another hand, the fact of favoring long documents relies on the assumption that long documents containing interesting passages may not be ranked at the top of the list since being likely to contain some supplementary information. Passage Retrieval approaches (see for example [8])

address this problem by ranking documents w.r.t. the similarity of their passages or thematic segments with the query. With such approaches, the fact that relevant information is surrounded by parts of text deviating from the interesting topic does not penalize the document containing it. Nevertheless, in the context of classical Information Retrieval, we believe that favoring long documents on the base they may contain hypothetical relevant information risks to induce some biases. This indeed may favor some long documents in which query’s terms figure but in scattered way, rendering not the presence of relevant information.

At last, the TREC conference uses a pooling method to determine relevant documents [10]. The first documents retrieved by each participant system are examined to assess their relevance potential. Some relevant documents may thus not have been identified. Zobel estimated the total number of relevant documents of the collection in [18], by extrapolating the difference between the numbers of relevant documents identified in the k and $k + 1$ first documents retrieved by the participant systems. If the participant systems of the pooling process favor the long documents, long relevant documents have more chances to be identified. This bias could somewhat explain the greater number of long relevant documents. Studying the distribution of relevant documents, in such a corpus, w.r.t. to their length to establish a new normalization function, as it is the case for the pivoted normalization, may lead to favor again slightly more long documents, which increases the bias existing in the corpus, and so on. Using the extrapolation method [18] over each bin separately may allow to solve this problem.

In a more general way, we believe that the fact of favoring long documents is not a good solution to retrieve useful documents as reponse to a user’s query. Nevertheless, measures adopting such an approach are likely to obtain better effectiveness results. We thus propose to substitute a notion of precision of retrieved terms $TPrec$ for the precision of retrieved documents $Prec$ usually used. The term precision computes the ratio of terms belonging to relevant documents within the set of terms contained by the n first retrieved documents. Using this new precision criterion may reduce the better evaluation results obtained by systems favoring long documents. It better renders the ratio of gain/cost, the interesting information collected w.r.t. the effort furnished by the user to find it. The effectiveness of each method is reevaluated in Table 4 by using the mean average term precision, $MATP$, and the term precision at rank 100, $TP@100$, computed same manner as in Section 3 but using this new term precision $TPrec$. This measure limits the bias induced by the fact that longer documents are more likely to be relevant. For example, better results obtained by Pivoted Smart are really minored when using the term precision criterion. However, measures that favor long documents being likely to retrieve more and longer relevant

Smart	Title		Narrative	
	MATP	TP@100	MATP	TP@100
ZIFF	0.210	0.149	0.379	0.264
AP	0.176	0.121	0.355	0.227
WSJ	0.203	0.190	0.429	0.321
FR	0.175	0.087	0.384	0.168

Okapi	Title		Narrative	
	MATP	TP@100	MATP	TP@100
ZIFF	0.225	0.179	0.347	0.292
AP	0.182	0.126	0.335	0.214
WSJ	0.239	0.190	0.417	0.286
FR	0.174	0.080	0.331	0.125

Inquery	Title		Narrative	
	MATP	TP@100	MATP	TP@100
ZIFF	0.124	0.117	0.081	0.104
AP	0.158	0.105	0.237	0.143
WSJ	0.184	0.147	0.221	0.142
FR	0.181	0.070	0.068	0.034

Pivoted Smart	Title		Narrative	
	MATP	TP@100	MATP	TP@100
ZIFF	0.138	0.153	0.294	0.204
AP	0.172	0.116	0.346	0.216
WSJ	0.220	0.157	0.406	0.268
FR	0.116	0.042	0.299	0.074

Table 4. Term precision results

documents, such measures may still be slightly favored.

6 A Study of the Similarity Scores

We propose, in this section, to study the similarity scores of the measures rather than the probabilities of retrieval of the documents. Our purpose is not only to study the impact of the document length on the mathematical expectation of similarity but to assess the influence of the query size too. So, we cannot use the restricted set of queries of TREC. Therefore, we preferred to study the similarities of documents with artificial queries in order to work with a sufficiently significant and heterogeneous set. For each query size between 1 and 200 terms varying by a step of 5 terms, queries are produced by taking each term for a same query in a randomly chosen document containing more than 300 words⁴. We focus on the average of the similarity scores of documents of a same bin, containing at least one term in common with the query⁵, w.r.t. different sizes of query. The average similarity of documents of a same bin is computed for each produced query. Figure 2 plots the mean similarity for each bin (of documents of the ZIFF corpus) and four query sizes QL (1, 10, 50 and 200 meaningful terms) w.r.t. the three measures presented in Section 3⁶, each point corresponding to the average of scores obtained over 1000 artificial queries.

We first remark that no measure is fair w.r.t. document and query length. The query size is very influent on the expectation of document similarity. Indeed, for the Smart and Okapi measures, short documents appear to obtain a better average similarity score than long ones with short queries, whereas with longer queries, the long documents appear to have a great advantage. We could explain this inversion of dominance by the fact that the advantage of short docu-

⁴Using only terms co-occurring in a document enables to produce relatively coherent queries. Used documents contain more than 300 words in order to produce queries with a sufficient variety of terms.

⁵Documents having no term in common with the query have no chance to be ranked at the top of the list and are thus not concerned by the normalization. The fact that the number of short documents containing no query’s terms is greater may bias the experiments if taken into account.

⁶Scores of Inquery and Okapi have been divided by 100 to be plotted on the same graph than Smart scores.

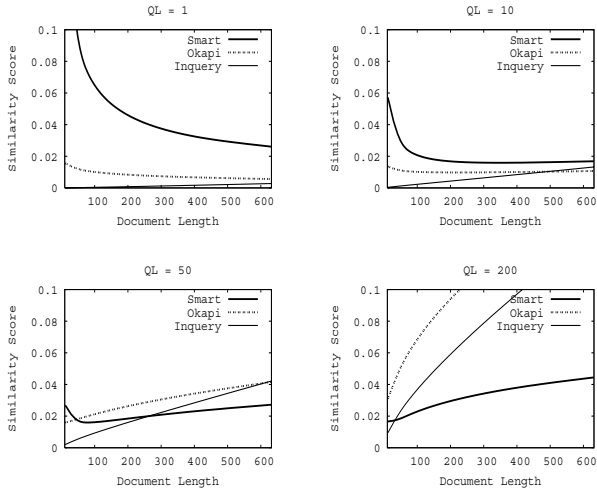


Figure 2. Similarity score tendencies

ments, i.e., a lower probability to contain terms figuring not in the query, is exceeded by the advantage of longer ones, i.e., a greater probability to contain query's terms, when the query size increases. With the third measure, i.e., Inquiry, the advantage of long documents is already greater when the query contains few terms and is intensified when the query size increases. Assuming that the expectation of similarity is correlated to the retrieval probability, the observation made in [4], i.e., that the slope of the pivoted normalization should depend on the query size, appears to be confirmed. However, given the shape of the curves, it is clear that the normalization cannot effectively be achieved by a linear pivoting. Moreover, with extremely long queries, the retrieval probability of longest documents exceeds their relevance one. Contrarily to what is done in [4], their similarity score should then be decreased.

A study of the maximal similarity score obtained within each bin with each similarity measure for each produced query has been realized same manner. The mean maximal score curves follow roughly the same variations than the mean average similarity ones given in Figure 2. The scaling of the similarity scores appears then to be constant. A modelization of the mean similarity $Mean(DL, QL)$ w.r.t. the document and query lengths, DL and QL , could allow us to determine a normalization of the similarity scores:

$$NormSim(D, Q) = \frac{1 + Sim(D, Q) - Mean(DL, QL)}{2} \quad (3)$$

Such a normalization would enable to adjust the similarity of a document w.r.t. to its mathematical expectation rather than simply give more weight to the long documents as it is the case in the pivoted normalization.

7 Normalization by Statistical Regression

This section aims to define the $Mean(DL, QL)$ function mentioned above. This search of function is realized by a statistical regression⁷ of the distribution of similarities obtained in the experiments conducted in Section 6 w.r.t. the Smart measure. After having determined variables and coefficients of the model, we experiment its validity and its impact on Information Retrieval results.

Statistical regression [17] is a way of describing how one variable, the outcome, is numerically related to predictor variables. A regression equation allows to express the relationship between variables algebraically by means of equation of the following type:

$$y = a_0 + a_1 \times x_1 + a_2 \times x_2 + \dots + a_p \times x_p \quad (4)$$

where, y is the variable to be predicted, the outcome, $x_1 \dots x_p$ are the predictor variables and $a_0 \dots a_p$ are the parameters, coefficients, of the equation.

The first step of the regression process is to determine the variables of the model. In our case, the two lengths cannot be considered independently as two predictor variables since, given the results obtained previously, the document length has not the same impact w.r.t. the different query sizes. Face to the difficulty of finding directly a modelization of the function w.r.t. the document and query sizes, we have chosen to work with results obtained with the different query sizes independently. So, we first aimed to obtain a set of equations, one for each query size studied, modeling the expectation of similarity w.r.t. the document length.

The search of the best variables is realized by the maximization of a correlation coefficient $R_{y, x_1 x_2 \dots x_p}^2$, which renders the strength of the relationships of the p predictors with the outcome y , within the whole set of N observations [17]. With two predictor variables, it is computed as:

$$R_{y, x_1 x_2}^2 = \frac{R_{y, x_1}^2 + R_{y, x_2}^2 - 2R_{y, x_1} R_{y, x_2} R_{x_1, x_2}}{1 - R_{x_1, x_2}^2} \quad (5)$$

where, $R_{a, b}$ is the correlation coefficient of the two variables a and b , in fact the covariance of a and b divided by the variance of a times the one of b .

Several sets of variables having been studied over the four corpuses, we have selected the model that owns the best correlation coefficients. It corresponds to the following equation for a given query size QL , DL and QL corresponding to the number of unique meaningful terms con-

⁷Statistical regression in the field of Information Retrieval has already been experimented in [5] but the point of view is different: contrarily to our approach that aims to normalize existing measures w.r.t. text sizes, the goal was to determine a new similarity function based on a learning of relevance judgements, what correspond to measures, criticized in Section 5, which adapt their retrieval probability to the relevance one.

tained in the document and the query respectively⁸:

$$Mean_{QL}(DL) = a_0 + a_1 \times \ln(DL) + a_2 \times \ln(\ln(DL) + 1) \quad (6)$$

Variables being determined, the point is to determine the coefficients a_0 , a_1 and a_2 . This is done by the least square method [3] which assumes that the best-fit curve is the one having the minimal sum of squared deviations (least square error) from a set of observed datas. In our problem, it comes down to resolve the following system:

$$\begin{cases} a_0 \times n + a_1 \times \sum_{i=1}^n x_{1,i} + a_2 \times \sum_{i=1}^n x_{2,i} = \sum_{i=1}^n y_i \\ a_0 \times \sum_{i=1}^n x_{1,i} + a_1 \times \sum_{i=1}^n x_{1,i}^2 + a_2 \times \sum_{i=1}^n x_{1,i}x_{2,i} = \sum_{i=1}^n x_{1,i}y_i \\ a_0 \times \sum_{i=1}^n x_{2,i} + a_1 \times \sum_{i=1}^n x_{1,i}x_{2,i} + a_2 \times \sum_{i=1}^n x_{2,i}^2 = \sum_{i=1}^n x_{2,i}y_i \end{cases} \quad (7)$$

where n is the number of observations used for the training of the parameters, y_i the i th outcome observed and $x_{1,i}$ and $x_{2,i}$ the values of both variables when considering the i th observation. Figure 3 plots the parameters obtained over

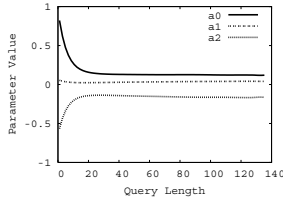


Figure 3. Parameter values

the ZIFF corpus for the different $Mean_{QL}(DL)$ functions w.r.t. the query size used. A modelization of the evolution of these three coefficients has been realized to establish the final $Mean(DL, QL)$ function that combines these multiple functions $Mean_{QL}(DL)$. Using the whole set of $Mean_{QL}(DL)$ functions (41 functions for the 41 query sizes studied between 1 and 201 meaningful terms), the regression leads to the following equation:

$$\begin{aligned} Mean(DL, QL) = & (1.00586 + 0.18685 \times \ln(QL) - 1.02757 \times \ln(\ln(QL) + 1)) \\ & + \ln(DL) \times \\ & (0.09036 + 0.02671 \times \ln(QL) - 0.10388 \times \ln(\ln(QL) + 1)) \\ & + \ln(\ln(DL) + 1) \times \\ & (-0.77143 - 0.16194 \times \ln(QL) + 0.803 \times \ln(\ln(QL) + 1)) \end{aligned} \quad (8)$$

Whereas the coefficient $R_{y,x_1x_2}^2$ evaluates the relationships between the variables, the coefficient $R_{y,\hat{y}}^2$ computes the correlation between observed and predicted values:

$$R_{y,\hat{y}}^2 = 1 - \frac{Non\ Explicated\ Variability}{Total\ Variability} = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (9)$$

⁸Experiments realized by using the total number of meaningful terms have shown that the number of unique meaningful terms is better correlated to the outcome.

where, N is the total number of observations realized and \bar{y} the average of the outcomes of this set. Comparing calculated values with real ones, this coefficient includes the parameters of the model in its evaluation. It then allows us to assess the representativity of the training set. In the aim to determine the size of the set of observations needed to effectively train the parameters of the model, different subsets of the whole set of observations have been experimented. In this way, we note $T(c, x)$ a training set of observations realized on each bin of a corpus c for the x first query sizes. On an other hand, the $R_{y,\hat{y}}^2$ coefficient allows us to assess the portability of parameters trained on one corpus to others.

$R_{y,\hat{y}}^2$	$Mean(DL, QL)$				$R_{y,\hat{y}}^2$	$Mean(DL, QL)$			
	ZIFF	AP	WSJ	FR		ZIFF	AP	WSJ	FR
$T(ZIFF, 5)$	0.938	0.931	0.909	0.719	$T(WSJ, 5)$	0.948	0.949	0.931	0.859
$T(ZIFF, 10)$	0.981	0.981	0.971	0.89	$T(WSJ, 10)$	0.976	0.969	0.975	0.909
$T(ZIFF, 20)$	0.992	0.984	0.984	0.937	$T(WSJ, 20)$	0.987	0.966	0.992	0.946
$T(ZIFF, 41)$	0.994	0.976	0.99	0.959	$T(WSJ, 41)$	0.988	0.948	0.995	0.968
$T(AP, 5)$	0.923	0.915	0.892	0.661	$T(FR, 5)$	0.948	0.931	0.953	0.909
$T(AP, 10)$	0.954	0.95	0.93	0.756	$T(FR, 10)$	0.971	0.929	0.967	0.944
$T(AP, 20)$	0.974	0.98	0.958	0.845	$T(FR, 20)$	0.977	0.897	0.951	0.982
$T(AP, 41)$	0.987	0.993	0.977	0.913	$T(FR, 41)$	0.972	0.872	0.934	0.972

Table 5. Coefficients of correlation

A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. Results of Table 5 show that the models are relatively robust since parameters computed w.r.t. any whole set of observations appear to be well suited over each other corpus. Nevertheless, the ZIFF corpus owning a greater variety of document lengths (mean length of bins comprised between 20 and 1600 unique terms), the model trained on this corpus appears to be the most portable. On the other hand, the training set does not need to contain more than 20 query sizes, $R_{y,\hat{y}}^2$ being roughly the same for larger subsets of observations. Therefore, if a training is even needed, its cost is not very expensive.

Table 6 gives Information Retrieval results obtained by the Smart measure normalized by statistical regression RSmart, using Equations 3 and 8. RSmart results are clearly

RSmart	Title				Narrative			
	MAP	P@100	MATP	TP@100	MAP	P@100	MATP	TP@100
ZIFF	0.192	0.146	0.225	0.180	0.306	0.207	0.371	0.271
AP	0.174	0.118	0.189	0.132	0.333	0.216	0.361	0.239
WSJ	0.202	0.159	0.235	0.183	0.367	0.272	0.429	0.337
FR	0.154	0.061	0.198	0.093	0.234	0.106	0.335	0.177

Table 6. Normalized Smart results

better than Smart ones when using the title of the queries. With the whole text of the query, long documents were favored by Smart and, given the fact that longer documents have more chances to be relevant, this measure obtained really good results. Nevertheless, due to the fact that our normalization allows relevant short documents to be retrieved, our results are better in term of term precision. As discussed

in Section 5, we believe that measures favoring long documents, assuming that they have more chances to be relevant, obtain good results for wrong reasons. The term precision criterion limits this bias but we are convinced by the fact that our results, already better than ones obtained by other measures, would dominate them really more over corpuses owning a uniform distribution of relevant documents.

An additional Student t-test⁹ has shown that the differences between our results and the best results obtained by other measures are statistically significant with a 99% confidence rate, all values being greater than the p-value 2.57.

8 Application to Inter-Text Similarities

Similarity measures, being not only employed in the field of Information Retrieval, are also useful to compute the proximity of documents or parts of texts, for example in the fields of document clustering [15] or thematic text segmentation [9]. Nevertheless, whereas the document length normalization has widely been studied to compute the similarity of documents with queries, it has, up to our knowledge, not really been experimented when applied to inter-text similarities yet. We then assess the validity of our normalization when computing the similarity of two texts and then study its impact on a clustering process of documents.

Smart being devoted to retrieve documents as response to queries, its term weights have to be adapted to compute similarities between documents. In that way, the similarity $Sim(DL, D'L)$ uses identical term weights for both documents DL and $D'L$: $w_{D_i} = w_{D'_i} = (1 + \log(tf_i)) \times \log \frac{N}{n_i}$. The statistical regression of the similarity distribution of this new measure, realized same manner as in Sections 6 and 7, leads to the following equation:

$$\begin{aligned} Mean(DL, D'L) = & (0.68296 + 0.13079 \times \ln(D'L) - 0.70593 \times \ln(\ln(D'L) + 1)) \\ & + \ln(DL) \times \\ & (0.05654 + 0.0164 \times \ln(D'L) - 0.06364 \times \ln(\ln(D'L) + 1)) \\ & + \ln(\ln(DL) + 1) \times \\ & (-0.50831 - 0.10729 \times \ln(D'L) + 0.52842 \times \ln(\ln(D'L) + 1)) \end{aligned} \quad (10)$$

The point is now to determine whether this equation fits the distribution of similarities between real documents. Then, Figure 4 plots the mean similarity between documents of different couples of bins of the corpus ZIFF. The correlation coefficient $R^2_{y,\hat{y}}$ between estimated and observed values is equal to 0.92. Equation 10 appears thus to be well suited to describe inter-documents similarities. Nevertheless, due to the training sets imperfections, $Mean(DL, D'L) \neq Mean(D'L, DL)$ contrarily to what should have been observed. To overcome this problem, the normalization func-

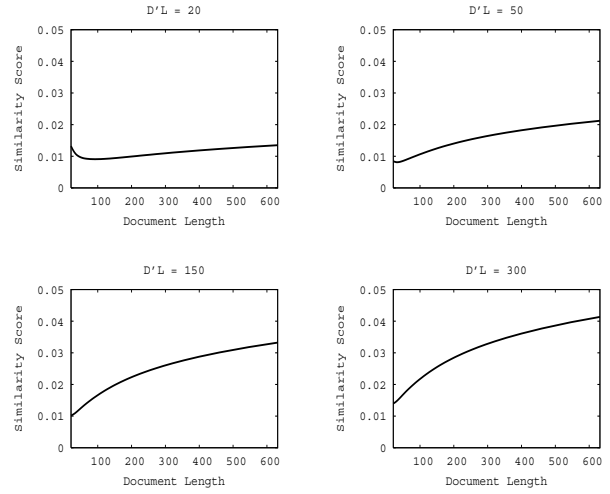


Figure 4. Documents similarity tendencies

tion (Equation 3) becomes:

$$NormSim(D, D') = \frac{1}{2} \times (1 + Sim(D, D') - (Mean(D_L, D'_L) + Mean(D'_L, D_L))/2) \quad (11)$$

Similarity expectation variations may lead a clustering process to favor some groupings of documents only because of their size. Therefore, a similarity normalization may increase the clusters quality. Similarity scores being changed by the normalization technique, it is not possible to assess the clusters quality in term of intra and inter clusters measures. Therefore, we chose to assess the impact of the normalization over the cluster hypothesis [15]. This hypothesis states that relevant documents in response to user queries tend to be more related than others and thus that the application of a clustering technique on the first retrieved documents by a classical Information Retrieval system may enable to obtain a cluster containing a great proportion of relevant documents. In [15], authors claimed that the best suited clustering algorithms are the hierarchic agglomerative ones and that the best results may be obtained with the Group Average Algorithm, where the distance between two clusters depends on the average distance between their respective documents. This algorithm has been run over the 100 first retrieved documents by Smart as response to each Title query over each corpus. Clusters are evaluated by a measure E combining recall (Rec), proportion of relevant documents that are contained by the cluster, and precision (Prec), proportion of cluster's documents that are relevant:

$$E = 1 - \frac{(\beta^2 + 1) \times Prec \times Rec}{(\beta^2 \times Prec) + Rec} \quad (12)$$

where β determines the relative importance of the recall w.r.t. the precision, $\beta = 1$ allocating equal importance to both measures, $\beta = 2$ attributing twice importance to recall

⁹The 99% Student t-test is based on the average, the standard deviation and the cardinal of a set of runs. It aims to insure with a rate of confidence of 99% that the difference in means of two sets is significant.

w.r.t. precision and $\beta = 0.5$ twice importance to precision w.r.t. recall. The effectiveness of the clustering process is evaluated by an ‘optimal cluster search’ [15], the search of the best cluster that can be reached in the hierarchy w.r.t. a specific query, which presents the advantage that it isolates the cluster effectiveness from the bias introduced by the search method employed. Table 7 reports then the best (the least) E scores (w.r.t. β) existing in the hierarchies built w.r.t. original similarities (Sim) or normalized ones (NormSim). According to the results of Table 7, the normalization appears to really increase the quality of the produced clusters, since each score obtained with the normalized measure is significantly lower than the corresponding one obtained with the original measure.

Group	Sim			NormSim		
	$\beta = 0.5$	$\beta = 1$	$\beta = 2$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$
Average	0.5511	0.6233	0.6324	0.5312	0.5908	0.6075
ZIFF	0.5511	0.6233	0.6324	0.5312	0.5908	0.6075
AP	0.5480	0.6280	0.6291	0.5264	0.6049	0.6124
WSJ	0.5966	0.6860	0.7005	0.5802	0.6527	0.6784
FR	0.4258	0.4932	0.5210	0.4203	0.4898	0.5156

Table 7. Quality of the optimal cluster

9 Conclusion

In this paper, we presented a new document length normalization technique that relies on a statistical regression of the similarity expectation w.r.t. documents and queries sizes. Several existing normalization techniques assume that long documents are more likely to be relevant and have to be favored. We believe that such approaches, in spite of their good results, are not good solutions for the Information Retrieval problem, since short relevant documents are really penalized whereas they may be at least as useful as long ones. Have writers to produce documents as long as possible to be read? Consequently to these observations, our normalization aims to give the same similarity expectation for each document and query size. The results obtained in Section 7 show that our approach is promising.

The normalization has been finally applied to inter-document similarities, in particular in the field of document clustering. Using classical measures, some documents have more chances to be grouped since their sizes lead to a better expectation of similarity. The normalization of these similarities appears to allow the clustering method employed in Section 8 to produce more interesting groups.

As future works, we plan first to model the similarity expectation of Okapi and Inquery as it is the case for Smart in this paper, the supplementary informations these measures use may lead to better results. Additionally, several variables, as the maximal tf or the normalization factor of Smart, can be added to the models in the aim to better describe the similarity expectation between a document and a

query. At last, we have to assess the impact of this normalization on every applications using text similarities, such as text segmentation or passage retrieval for examples.

10 Acknowledgments

This work was supported by “Angers Loire Metropole”.

References

- [1] J. Allan, J. P. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. C. Swan, and J. Xu. INQUERY does battle with TREC-6. In *TREC*, pages 169–206, 1997.
- [2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press/Addison-Wesley, 1999.
- [3] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.
- [4] T. L. Chung, R. W. P. Luk, K. F. Wong, K. L. Kwok, and D. L. Lee. Adapting pivoted document-length normalization for query size: Experiments in chinese and english. *ACM TALIP’06*, 5(3):245–263, 2006.
- [5] W. S. Cooper, F. C. Gey, and A. Chen. Probabilistic retrieval in the tipster collections: An application of staged logistic regression. In *TREC*, pages 73–88, 1992.
- [6] D. Harman. TREC-3. *SIGIR Forum*, 27(3):19–23, 1993.
- [7] P. S. Jacobs. Introduction: Text power and intelligent systems. In P. S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pages 1–8. Erlbaum, Hillsdale, 1992.
- [8] M. Kaszkiel and J. Zobel. Effective ranking with arbitrary passages. *Journal of the American Society of Information Science*, 52(4):344–364, 2001.
- [9] S. Lamprier, T. Amghar, B. Levrat, and F. Saubion. Seggen: A genetic algorithm for linear text segmentation. In M. M. Veloso, editor, *IJCAI*, pages 1647–1652, 2007.
- [10] C. D. Loupy and P. Bellot. Evaluation of document retrieval systems and query difficulty. In *LREC’00*, 2000.
- [11] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [12] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *TREC*, pages 21–30, 1992.
- [13] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Proces. Manage.*, 24(5):513–523, 1988.
- [14] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Inf. Proces. Manage.*, 32(5):619–633, 1996.
- [15] A. Tombros, R. Villa, and C. J. V. Rijsbergen. The effectiveness of query-specific hierarchic clustering in information retrieval. *Inf. Proces. Manage.*, 38(4):559–582, 2002.
- [16] E. M. Voorhees and D. Harman. Overview of the fifth text retrieval conference (trec-5). In *TREC*, 1996.
- [17] G. U. Yule. On the theory of correlation. *Journal of the Royal Statistical Society*, 60(4):812–854, December 1897.
- [18] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Research and Development in Information Retrieval*, pages 307–314, 1998.
- [19] J. Zobel and A. Moffat. Exploring the similarity space. *SIGIR Forum*, 32(1):18–34, 1998.