Improving DFR Models for Information Retrieval in Cognitive Contexts by Adjusting the Documents Length Normalization

Eldon Caldwell Marín and Mauricio Zamora
University of Costa Rica
Costa Rica

Antonio Ferrández Rodríguez
University of Alicante
Spain

Abstract

Cognitive contexts such as customer service of semantic web systems need autonomous information retrieval processes. Deviation from randomness (DFR) is a methodology for modeling unstructured information retrieval using a weighting to generate a relevance ranking of documents based on the concepts of information content and information gain. This approach is very effective compared with others such as the Okapi-SLM model, however, proposes a document length normalization that uses the simple average of these lengths throughout the collection which may introduce statistical bias. Through quasi-experimental methodology we compare DFR basic model with an approach we have called DFR_{adj} introducing an adjustment in the second normalization with the average lengths of the documents in the collection containing the search terms. As a result, it was found that the proposed adjustment allows an incremental improvement over the effectiveness of DFR basic model and it opens an interesting research line exploring the effect of new normalizations when the variability of length documents is high. We contribute to the state of the art finding evidence related to the influence of the documents length in the collection and specifically those containing the search terms; and we conclude this is crucial to achieving better heuristic solutions.

Keywords
Information retrieval, deviation from randomness, natural language processing, probabilistic models for information retrieval.

1. Introduction

With the advent of the internet and information sharing networks in real-time, effective information retrieval has become a critical need. Particularly, it is relevant when the information is stored in an unstructured way (Amati G. and Van Rijsbergen C.J., 2002).

Today, research on the development of better ways of retrieving information is geared to various problems in cognitive contexts, among which is, for example, the customer service of semantic web systems, the provision of semantic robotic devices, the design of robust machine learning algorithms, smart manufacturing systems, the development of cognitive factories and cognitive value chains (Carpineto, C. et al, 2012).

The unstructured information retrieval systems (RI) select and retrieve documents that are relevant for users, according to previously submitted information needs. As a result, these systems return documents ordered according to cutoffs that determine the correspondence that has the information in the document and the need expressed by the user (Amati G, 2003).

One of the fundamental principles of IR systems is that not all the relevant words have the same discriminatory value. Therefore, various techniques have been developed to calculate and assign weights to words according to their discriminatory power (Ponte J.and Croft W. B., 1998). A successful approach is the information retrieval process based on probabilistic models and the DFR model (Deviation from Radomness), published in 2002, has become an obligatory reference (Amati G. et al, 2002).

This paper presents an overview of the DFR methodology and a quick description of its criteria basis. Then we make a proposal with an statistical adjustment to the model and present the results of a controlled quasi-experiment comparing the new procedure and according to standard indexing techniques, thus for conclusions and future research lines.

2. Literature Review

A classic information retrieval system is shown in Figure 1. Users enter information need to be retrieved from a collection of unstructured documents. It performs a process of consultation and the RI system scans the entire collection and returns a set of relevant documents from an algorithm or discriminatory rule (Vilares, 2008).

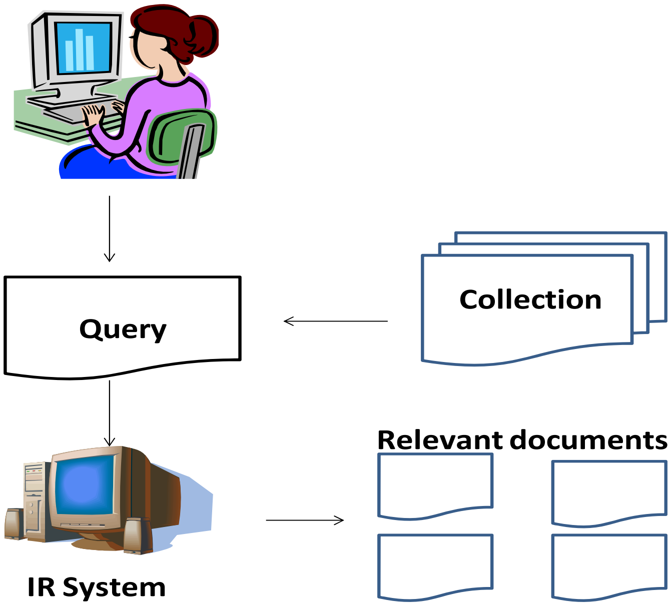

Figure 1: Information retrieval system in a basic model (Vilares, 2008).

The initial approach DFR probabilistic model is identical to other classical models of Information Retrieval (IR). As with any IR system, it starts with a need for information entered by the user in a query and a collection of documents indexed on which the search is performed (Amati G. et al, 2001). In a cognitive robotics application, this needs come from the immediate environment or human interaction, for example, instructions translated through voice recognition (Carpineto, Romano et al, 2013).

From internal representations typically in the form of text, the system tries to identify those documents that meet the need, ie which documents are relevant to the query.

The difference lies in how and based on what correspondence is calculated between the query and the documents in the collection. For example, in the case of vector model (Baeza-Yates and Ribeiro-Neto, 1999; Robertson S., Walker S., and Hancock-Beaulieu M., 1998), it is calculated with a formal mathematical basis (vector algebra), but there is no theoretical result to suggest that the process for calculating these correspondences is the correct or most appropriate.

Although it uses a mathematical basis, in a way, this approach passes blindly making changes and approximations in the calculation scheme of correspondences and, then, experimentally verify if the assumptions were correct and could improve the results.

Given a query, any IR system until today, reaches an uncertain understanding of the need for information that the query represents. Similarly, given the representations of the query and the documents, these systems currently can only try, with a level of uncertainty, the provision of a determination about whether the content of a document responds to the need of information or if it does.

The IR algorithms based on probabilistic models, initially raised in the '70s, resurfaced hard from the 90s, currently enjoying much attention from researchers in IR (Robertson, 1977). Under this approach, the mathematical functions that analyze the content matching the needs estimate the probability that a document is relevant to the query, as opposed to the concept of degree or measure of relevance of other models (Robertson and Belkin, 1978; Zhai and Lafferty, 2001). Consequently, the returned documents can be sorted based on the estimated probability of relevance to the query, instead of using a similarity measure.

The principle of sequencing by chance (probability ranking principle) is the theoretical basis on which probabilistic models are based. This principle shows that optimum recovery is one in which the documents are returned sorted in descending order according to their probability of relevance to the query terms.

More specifically, a probability model normally returns documents ranked according to the probability **P (R|dj,q)** of a document **dj** belongs to the set **R** of relevant documents to a query **q**, in other words, the probability that a document **dj** is relevant to a query **q** (Vilares, 2005).

Moreover, if in addition to an optimal sequence is also desired optimal result set, it is possible to apply Bayes' theorem, according to which the documents returned-those considered relevant-should be those for which the probability **P (R|dj,q)** to be relevant to the query is greater than the probability **1-P (R|dj,q)** not to be relevant to the query. That is, according to this principle, a document **dj** is relevant to a query **q** if and only if:

P(R|dj,q) > 1-P (R|dj,q)

DFR, rather than a model, is a methodology for modeling recovery, which allows the assignment of weights to generate the ranking of relevant documents that are related to the user's information need (Carpineto, Kuznetsov, et al, 2013). Like others, starts from the assumption that if a word appears in a document much more than expected, this document addresses this issue.

For practical purposes, in the following sections of this document will be used basic variables below (Amati et al, 2007):

ft : the number of occurrences of the term t in the collection.
ft,d: the number of occurrences of the term t in the document d.
ft,q: the number of occurrences of the search term t on query q.
nt: the number of documents in which t occurs.
D: the number of documents in the collection.
T: the number of terms in the collection.
λt: the ratio between ft and T.
ld: the length of the document d.
lq: search length q
avr ld: average length of documents in the collection.

The DFR methodology, whose main exponent is the system Terrier (Amati et al, 2007), is based on Poisson model-2 defined by Harter in 1975. However, houses important differences from other popular approaches such as the Okapi probabilistic model (Robertson S. et al, 1994; Robertson S., Walker et al, 1998).

First, DFR models are not based on the principle of sorting by chance, since they do not work on the concept of probability of relevance of a document, but based on the concepts of information content and gain information (Amati and Van Rijsbergen, 2002). Thus, the documents are not sorted according to their probability of relevance to the query, but regarding the information gain obtained, besides returning weight concept **wij** of a term in a document **dj** to the calculation of a gain term level.

Second, the DFR are nonparametric models, so, they don´t need any experimental adjustment for parameters, as in the case of Okapi-SLM model (Carpineto C. et al 2001; Amati G. et al, 2003). In the DFR logic, if one assumes that the distribution of terms in documents throughout the collection should be random (according to the Poisson model-2), then it is possible to measure the amount of information carried by a term in a document **dj** based on the difference between the actual distribution of the document and its expected distribution under the random model (Spärck Jones K. et al, 2000).

That is, if a word appears in a document many more times than could be estimated probabilistically, then it seems logical to assume that this document is on the subject.

When we are defining a DFR model, we must to determine its three components, which are specified in the calculation of the weight **wij** of a term in a document **dj**:

$$\mathbf{w_{ij} = tf_{iq} \times Inf_1\ (tf_{ij}) \times Prisk\ (tfn_{ij})}$$

where,

tf_{iq}: the term frequency in the query q
Inf_1: the informational content of the term **ti** in the document **d_j**.
Prisk: an expression of the risk assumed by accepting the term **t_i** as a valid descriptor of the document **d_j**.
tf n_{ij}: the frequency **tf_{ij}** of the term **t_i** in the document **d_j** after being normalized based on the length of the document.

Any DFR model has three components (Carpineto et al, 2001). The first one is the randomness model whereby it is assumed that the terms are distributed and which is given by a probability function **$Prob_1$** where **$Prob_1$** (**tf_{ij}**) is the probability that the term **tf** appears **tf_{ij}** times in document **dj**. One of the simplest cases is to use a binomial distribution as follow (but there is another options, for example, to use the geometric distribution):

$$\mathbf{Prob_1\ (tf_{ij}) = \binom{TFi}{tfij}\ X\ p^{tfij}\ X\ q^{Tfij\ -\ tfij}}$$

where,

$$\mathbf{p = 1/N \text{ and } q = 1\text{-}p}$$

tfij is the term frequency ti in the document dj
TF_i is the total frequency of the term **t_i** in the collection
N is the number of documents in the collection

According **$Prob_1$** function is used, a different model is obtained with this calculation, and we can estimate the informative content of a term in a document (Levow G., 2005; Vilares J., 2008).

Collections have two types of words. First, the "specialty words" which are those of higher information content and are concentrated in the documents "elite" and, therefore, they are more useful for the recovery process and their distribution is different than expected from the random model. Second, there are the non-specialty words; those with a low informative content (as stop words), and which involves a random distribution along the collection (Savoy J., 2001).

So, if according to the distribution model adopted, a term has a high probability of appearing in a document, it is assumed that it is a word "non-specialty", ie low information content. Conversely, if a term has a low probability of appearing in a document, then it is a "word of specialty".

Based on these criteria, we define the informative content of a term in a document (**Inf_1**) as:

$$\mathbf{Inf_1 = -log_2\, Prob_1}$$

The second component of a DFR model is called “first normalization”. The basic idea behind this technique is:

"If an uncommon term (word of specialty), does not appear in a document, we can assume that the probability of being informative on the issue of the document is low or absent. Thus, if we accept that term as a valid descriptor of that document, we will be taking a very high risk because there are indicators that suggest that it is unreliable. Conversely, if an uncommon term appears repeatedly in a document, then we can assume that their probability of being informative on the issue of the document, is very high. Consequently, if we take it as a descriptor, the risk assumed in this case is very low... " (Vilares, 2008).

If Prob2 is the probability of a term ti to be informative in relation to the subject in the document dj, we can define the risk function Prisk associated to the fact of taking ti as term representative of **dj** as:

$$\mathbf{P_{risk} = 1 - Prob_2}$$

Normalization refers to this component is Inf1 multiplying by $\mathbf{P_{risk}}$ and so, weigh the initial informative content $\mathbf{Inf_1}$ of a term in a document based on the risk assumed by taking the term as a valid descriptor of that document.

You can mention two known models for the function $\mathbf{P_{risk}}$, one based on the Laplace law of succession, and is known as normalization L:

$$\mathbf{P_{risk}} = \frac{\mathbf{1}}{(\mathbf{tfij+1})}$$

and another based on binomial distributions, which is known as standardization B:

$$\mathbf{Prisk} = \frac{\mathbf{TFi+1}}{\mathbf{dfi\ X\ (tfij+1)}}$$

where **dfi** is the number of documents in the collection that contain the term **ti**.

The third component of a DFR model is called “second normalization” and pursues to normalize the term frequency of ti ($\mathbf{tf_{ij}}$) in the document **dj** based on the document length and the average length of a document in the collection. The resulting normalized value, $\mathbf{tf\ n_{ij}}$, is the value used when finally we calculate the weight, rather than the initial frequency $\mathbf{tf_{ij}}$. One way of performing this normalization is shown as follows:

$$\mathbf{tfn_{ij} = tf_{ij}\ X\ d\ l_{avg} / dl_j}$$

where $\mathbf{dl_j}$ is the document length and $\mathbf{d\ l_{avg}}$ is the average length of documents in the collection.

Another way of normalization is also referenced in comparison exercises (Amati et al, 2005) and is as follows:

$$\mathbf{tfn_{ij} = tf_{ij}\ X\ log_2\,(1+\,(d\ l_{avg} / dl_j))}$$

Then, the weight assigned will be:

$$\boldsymbol{\omega}_{\mathbf{t,d}} = [\mathbf{log_2\,(1+\lambda_t) + f^*_{t,d}\ X\ log_2\,((1+\lambda_t) / (\lambda_t))]\ x\ [\ (f_t + 1) / (n_t\ x\ (f^*_{t,d} + 1))\ ]}$$

where,

$$\mathbf{f^*_{t,d} = [\ ft\ X\ (log_2\,(1+(c\ x\ avr\_l_d) / l_d))\ ]}$$

where **c** is a free parameter which normally is automatically established (Ouni and He, 2005)

3. Exploring an adjustment over the second normalization of a DFR basic model

The second normalization, as mentioned in the previous section, takes as its starting point the length of document d and the average length of the documents in the collection. However, we think that the simple average, despite being in the presence of a huge number of documents, can introduce bias when many documents where the term does not appear also affect the weight assigned and there is an inherent variability in the lengths of the documents in the collection.

From this premise, we hypothesized that the approach of a DFR model can increase effectiveness if the second normalization is performed on the basis of the segment of the collection in which the term occurs.

An quasi-experiment was conducted using standard techniques of indexing a collection of documents in Spanish language, corresponding to the base EFE95 EFE94 and CLEF 2003 (questions 141-200) using the abbreviated statements (short questions). DFR was adjusted by changing the variable **dl_{avg}** for **dl_{avg} aj** referred to the average length of the documents in the collection, in which the term **t_i** appears; and not the size of the total collection assuming statistically large collections.

This is a quasi-experiment because we use specific collections and an specific language. It is not performed with a random assignment of searches and not implemented from multiple collections selected through a pattern of randomness. Therefore, the results can be considered valid for the specific context of application. However, the basis of documents used are experimentally validated over more than 15 years by the international scientific community, in this way, we can be satisfied if the new setting shows improved information retrieval because we can introduce strong assumptions of generalization. As Peters and Braschler (2003) say, “the popularity of the Internet and the consequent global availability of networked information sources for an increasingly vast public have led to a strong demand for efficient cross-language information retrieval (**CLIR**) systems that allow users to search document collections in multiple languages and retrieve relevant information in a form that is useful to them, even when they have little or no linguistic competence in the target languages” (Braschler and Peters, 2002). Therefore, CLIR systems allow users of internationally distributed knowledge bases to find and retrieve relevant information in whatever language it is stored.

The best known evaluation campaign for information retrieval systems is the **T**ext **RE**trieval **C**onference (**TREC**) series, organized in the United States since 1991, mostly by the National Institute for Standards and Technology (NIST) (Hiemstra and Kraaij, 1998). From 1997 to 1999, TREC included a track for the evaluation of Cross-Language IR for European languages (CLEF, 2011). This track was coordinated jointly by NIST and by a group of European volunteers that grew over the years. However, probably due to a lack of experience with **TREC**, the result was a set of very simplistic topics in all the languages, and there were also some problems with the translations because it is difficult for nonnative speakers to select the most appropriate and natural terms in a target language. Another problem was that the NIST assessors working in a foreign language needed much longer than normal to make the relevance judgments. These and other difficulties led to the decision to make native speakers responsible for topic preparation and relevance assessment in following campaigns.

The Cross-Language Evaluation Forum (**CLEF**) was launched in January 2000 with the goal of continuing and extending the activities begun in the **CLIR** track at **TREC**. The declared objectives of **CLEF** are declared as follow: to provide an infrastructure for the testing and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts; to construct test-suites of reusable data that can be employed by system developers for benchmarking purposes and, finally, to create an R&D (research and development) community in the cross-language information retrieval (**CLIR**) sector (CLEF, 2012).

The organization of the CLEF campaigns developed several tracks: multilingual information retrieval, bilingual information retrieval, monolingual (non-English) information retrieval, domain-specific retrieval, and interactive cross-language retrieval. Then, we used a valid CLEF collection in order to compare our results with the DFR basic model and the Okapi model too. These results and their analysis are presented in next section.

4. The quasi-experiment results

Table 1 shows the results obtained by performing the proposed quasi-experiment. As observed, we generate two comparisons. The first one concerns the increase or decrease in the effectiveness of the retrieval between Okapi-SLM and the proposed model. This is a recommended step because Okapi-SLM still provide a reference point to be overcome in the first instance. The second comparison was made with basic DFR model showing, in this case, an effectiveness of 49.07%, using a recommended constant **c = 4** and short questions.

Table 1: Results of the quasi-experiment using CLEF 2003 Spanish corpus (EFE 94 and EFE 95)

	P_{avg} without the second normalization adjustment	P_{avg} with the proposed second normalization adjustment	Absolute Deviation	% of improvement
Okapi basic model	46.06%	49.45%	3.39%	7.36%
DFR basic model, c=4	49.07%		0.38%	0.77%

The result of the proposed adjustment to the second normalization, in terms of the relevance of the selected documents (compared with the correct answers given by CLEF) is 49.45%, which means an absolute increase of 0.38% compared to the relevance of the documents selected without the proposed adjustment. This increase could be considered minimal, but represents an increase of 0.77% over the original DFR, which is a significant result, if one takes into account that have not been used NLP techniques or methods based on recursive algorithms.

NLP techniques allow us to explore the corpus under the criteria of syntactic and lexical use, which introduces the opportunity for improvement from the perspective of intrinsic relevance of each term in a specific language. In this case, the second normalization was not performed considering the documents in the collection that contains only the most relevant terms but all of them, which can be a critical to our quasi-experiment. But our justification is based on the initial research target: the search for evidence of improvement without changing the theoretical approach of the models selected for comparison.

Regarding Okapi which it is a typical criteria for comparison (Amati et al, 2003), the improvement is more significant (7.36%) and greater than the improvement achieved by DFR basic model without adjustment (6.53%).

5.Conclusions and Future Work

After the execution of the quasi-experiment, we conclude that the proposed adjustment to the second normalization of DFR basic model improves information retrieval, but effectiveness is not very strong. However, we can say that it does establish a relevant sign for model design efforts. That is, we think that the quasi-experiment gives us an optimization path because the adjustment indicates an incremental improvement in the information retrieval compared to that afforded by the DFR basic model and without the use of Natural Language Processing techniques.

Moreover, we contribute to the state of the art finding evidence related to the influence of the documents length in the collection and specifically those containing the search terms. We think this is crucial to achieving better heuristic solutions.

Some future research lines could be oriented to the study of consistency of the proposed adjustment to the DFR basic model using other collections in other languages. Additionally, the study of NLP techniques influence on the proposed adjustment; the study of the effect of the variability of the length of documents contained into the indexed collection; the study of the use of NLP with expansion techniques using the new proposed adjustment; the study of the relation between effectiveness and the proposed adjustment using segmentation by paragraph in the documents of the collection and the proximity of the search terms, and, additionally, to study the possible improvement of relevance considering the lexical type of the term in the second normalization.

6.References

Amati G. Probability Models for Information Retrieval based on Divergence from Randomness Thesis of the degree of Doctor of Philosophy, Department of Computing Science University of Glasgow, 2003

Amati G., Carpineto C., and Romano G. Italian monolingual information retrieval with prosit. In Proceedings of CLEF (Cross Language Evaluation Forum 2002, pages 182–191, Rome, Italy, 2002.

Amati G., Carpineto C., and Romano G., Fub at trec-10 web track: A probabilistic framework for topic relevance term weighting. In Proceedings of the 10th TextREtrieval Conference (TREC-10), NIST Special Publication 500-250, pages 182–191, Gaithersburg, MD, USA, 2001.

Amati G., C. Carpineto, and G. Romano. Fub at trec-12: Robust and web track. In Proceedings of the 12th Text REtrieval Conference (TREC-10), Gaithersburg, MD, USA, 2003.

Amati, G., Carpineto, C., Romano, G. (Eds.). Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings. Lecture Notes in Computer Science Springer Volume 4425, 2007.

Amati G. and van Rijsbergen C. J. Probabilistic models of information retrieval based on measuring divergence from randomness. ACM Transactions on Information Systems, 20(4):357–389, 2002.

Braschler Martin and Peters Carol: The CLEF Campaigns: Evaluation of Cross-Language Information Retrieval Systems. CEPIS UPGRADE III (3), 78-81, 2002.

Carpineto C.,. De Mori R, Romano G., and Bigi B. An information theoretic approach to automatic query expansion. ACM Transactions on Information Systems, 19(1):1–27, 2001.

Carpineto, C., Kuznetsov, S., Napoli, A. (Eds.). Proceedings of the Workshop Formal Concept Analysis Meets Information Retrieval (FCAIR 2013), Moscow, March 24, 2013 CEUR-WS.org 2013 Vol-977, 2013.

Carpineto, C., Romano, G. Semantic Search Log k-Anonymization with Generalized k-Cores of Query Concept Graph. Proceedings of the 35th European Conference on Information Retrieval (ECIR 2013), Moscow pp. 110-121, ECIR 2013 Shared Best Paper Award, 2013.

Carpineto, C., D'Amico, M., Romano, G. Evaluating subtopic retrieval methods: Clustering versus diversification of search results Information Processing & Management. Vol. 48, Issue 2, pp. 358-373, 2012.

CLEF. Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics Third International Conference of the CLEF Initiative, CLEF 2012, Rome, Italy, September 17-20, 2012. Proceedings Series: Lecture Notes in Computer Science Vol. 7488 Editors: Tiziana Catarci, Pamela Forner, Djoerd Hiemstra, Anselmo Peñas and Giuseppe Santucci, 2012.

CLEF. Multilingual and Multimodal Information Access Evaluation Second International Conference of the Cross-Language Evaluation Forum, CLEF 2011, Amsterdam, The Netherlands, September 19-22, 2011. Proceedings Series: Lecture Notes in Computer Science Vol. 6941 Editors: Pamela Forner, Julio Gonzalo, Jaana Kekäläinen, Mounia Lalmas and Marteen de Rijke, 2011.

Hiemstra D. and Kraaij W. Twenty-one at trec-7: Ad hoc and cross-language track. In Proceedings of the 7th Text REtrieval Conference (TREC-7), NIST Special Publication 500-242, pages 227–238, Gaithersburg, MD, USA, 1998.

Levow Gina-Anne and Oard Douglas W. and Resnik Philip: Dictionary-based techniques for cross-language information retrieval. Information Processing and Management 41 (3) ; 523-547, 2005.

Ouni and He. Term Frequency Normalisation Tuning for BM25 and DFR model, Proceedings of ECIR'05, 2005.

Ponte J.and Croft W. B. A language modeling approach to information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Reasearch and Development in Information Retrieval, pages 275–281, 1998.

R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley and ACM Press, Harlow, England, 1999.

Robertson Stephen E., Walker Steve, Jones Susan, Hancock-Beaulieu Micheline and Gatford Mike. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference (TREC 1994). Gaithersburg, USA, 1994.

Robertson Stephen E., Walker Steve, and Hancock-Beaulieu Micheline. Okapi at TREC-7. In Proceedings of the Seventh Text REtrieval Conference. Gaithersburg, USA, 1998.

Robertson Stephen E. and Belkin N.J. Ranking in principle. Journal of Documentation, 34(2): 93-100, 1978.

Robertson Stephen E. The probability ranking principle in IR. Journal of Documentation, (33):126-148, 1977.

Robertson Stephen E., Walker Steve, and Beaulieu M. M. Okapi at trec-7: Automatic ad hoc, filtering, vlc, and interactive track. In Proceedings of the 7th Text REtrieval Conference (TREC-7), NIST Special Publication 500-242, pages 253–264, Gaithersburg, MD, USA, 1998.

Savoy J. Reports on clef-2001 experiments. In Working Notes of CLEF 2001, Darmstadt, Germany, 2001.

Spärck Jones Karen, Walker Steve, and Robertson Stephen E. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (parts 1 and 2). Information Processing and Management, 36(6):779-840, 2000.

Vilares, Jesús, El modelo probabilístico: Características y modelos derivados, Information and Documentation General Magazine, 18: 343-363, Universidade da Coruña, Spain, 2008.

Vilares Jesús. Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español, Ph.D. Thesis, Computing Science Department, Universidade da Coruña, Spain, 2005.

Zhai C. and Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 334–342, New Orleans, LA, USA, 2001.

Biography

Eldon Glen Caldwell Marín, full professor (Cathedraticus) at University of Costa Rica with over 20 years of teaching and research experience. Bachelor and Master of Industrial Engineering at University of Costa Rica, Dr. Caldwell earned a Masters degree in Operations Engineering at ITESM, Mexico, Master in Financial Analysis and M.Sc. in Service Marketing at Interamerican University of Puerto Rico. Also, he earned a Masters degree in Health Services Management, UNED, Costa Rica and finally a Ph.D. in Industrial Engineering at the University of Nevada-Autonomous University of Central America. Currently he is a doctoral researcher in Information Retrieval at the University of Alicante, Spain, and he is Academic Excellence Prized at the Ph.D. Program in Inclusive Education at University of Costa Rica, where he is a doctoral researcher too. His research interests include artificial cognition, information retrieval, machine learning, robotics and intelligent development of methodologies and algorithms for implementing lean systems. Contact him at eldon.caldwell@ucr.ac.cr or egcm@alu.ua.es.

Mauricio Zamora, teacher and assistant professor at University of Costa Rica with over 10 years of teaching and research experience. Bachelor of Science in Systems Engineering at International University of the Americas, he is a doctoral researcher in Information Retrieval at the University of Alicante, Spain and researcher in the Industrial Robotics Laboratory at University of Costa Rica. His research interests include information retrieval, machine learning, robotics and natural language processing. Contact him at mauricio.zamora@ucr.ac.cr.

Antonio Ferrández is a Full-time Lecturer at the Department of Software and Computing Systems in the University of Alicante (Spain). He obtained his Ph.D. in Computer Science from the University of Alicante (Spain). His research interests are: Natural Language Processing, Anaphora Resolution, Information Extraction, Information Retrieval and Question Answering. He has participated in numerous projects, agreements with private companies and public organizations related to his research topics. Finally, he has supervised PhD Thesis and participated in many papers in Journals and Conferences related to their research interests. Contact him at antonio@dlsi.ua.es.