An Ideal Context for Information Retrieval

Rabeb Mbarek, Mohamed Tmar
Multimedia Information systems and Advanced Computing Laboratory,
High Institute of Computer Science and Multimedia,
University of Sfax, Sfax, Tunisia
rabeb.hattab@gmail.com, mohamedtmar@yahoo.fr,
http://www.miracl.rnu.tn

Abstract

In general, document representation and ranking are dependent on context. In this paper we compute a context which gives the best ranking. This context is called *ideal context*.

1 Introduction

Information Retrieval (IR) deals with the retrieval of all and only the documents which contain information relevant to any information need expressed by any user's query. A system matching that definition exists in principle. In practice a system is unable to answer any query with all and only relevant information because of the unsolvable task of understanding both the relevant information enclosed in documents and the information need expressed through any query submitted by any user.

To refine the IR process, it is required to apply the Relevance Feedback (RF) technique. RF usually consists in extracting keywords from relevant judged documents and then of adding some terms to the initial query in order to express the user need in a more expressive way. It has been shown that RF is an effective strategy in IR [4].

Since information needs evolve with many variables like user, place, and time, relevance is context-dependent. Therefore, IR is also context-dependent which represents the high complexity of IR.

In this paper the Vector Space Model (VSM) is adopted as an infrastructure. It is introduced in [7] and [6]. A recent reconsideration of the geometry underlying IR, and indirectly of the VSM, was done in [3]. In VSM, documents and queries are modeled as elements of a vector space. This vector space is generated by a set of basis vectors that correspond to the index terms. Each document can be represented as a linear combination of these term vectors. One can find a nice and a short introduction of VSM in [1, Section 3].

According to [1], a context is modeled by a basis and its evolution is modeled by linear transformations from one base to another. The basic idea is that, first, a vector is generated by a basis just as an information object is generated within a context. Second, every vector can be generated by different bases and belongs to infinite subspaces; this is consistent with the fact that every information object is generated within different contexts. Finally and as a corollary, the subspace spanned by a basis contains all those vectors that describe information objects in the same context. Note that RF is an example of context change [1].

The determinant of a triangular matrix is the product of the diagonal elements, its inverse is a triangular matrix and the product of two triangular matrices is a triangular matrix too. These facts assure that the manipulation of triangular matrices is performed within the space of triangular matrices. Thus providing advantages at computational level when modeling contexts and describing context changes [1].

In general, document representation and ranking are dependent on context. In fact, if T_1 and T_2 are two context matrices which, respectively, generate documents d_1 and d_2 with the same coefficient a, and if a query is generated by an arbitrary context matrix U with coefficient b, then $a^T.(T_1^T.U).b \neq a^T.(T_2^T.U).b$. There exists a context that provides the best document ranking. This context is called *the ideal context*. In this paper we give an algorithm to compute this ideal context.

This paper is organized as follows. In Sections 2 we compute an ideal context. In Section 3 some experiments are reported to explore some of the potential of the proposed approach. Section 4 concludes.

2 Ideal context

In general, the vector x of the document x authored in its own context is generated by the base T. The latter is in its turn not necessarily equal to the base U that generates a query y or another document. Therefore, x is represented by $x = T.a$ whereas y is represented by $y = U.b$ where a and b are the coefficients used to combine the base vectors of T and U, respectively. If relevance is estimated by the usual inner product, then documents are ranked by $x^T.y = (T.a)^T.(U.b) = a^T.(T^T.U).b$.

2.1 Scenario

Let R be the set of relevant documents and S be the set of irrelevant documents. The ideal context C is a context which gives the best ranking that maximizes the inner product between query vector and relevant document vectors and minimizes the inner product between query vector and irrelevant document vectors. Then the context C maximizes the sum of inner product between query vector and relevant document vectors and minimizes the sum of inner product between query vector and irrelevant document vectors. So

$$C = arg \max_{B \in \mathbf{T}_n(\mathbb{R})} \sum_{d \in R} d^T.B^T.q = arg \max_{B \in \mathbf{T}_n(\mathbb{R})} (\sum_{d \in R} d)^T.B^T.q \quad (1)$$

and

$$C = arg \min_{B \in \mathbf{T}_n(\mathbb{R})} \sum_{d \in S} d^T.B^T.q = arg \min_{B \in \mathbf{T}_n(\mathbb{R})} (\sum_{d \in S} d)^T.B^T.q \quad (2)$$

which implies that

$$C = arg \max_{B \in \mathbf{T}_n(\mathbb{R})} \frac{(\sum_{d \in R} d)^T.B^T.q}{(\sum_{d \in S} d)^T.B^T.q} \quad (3)$$

Where $\mathbf{T}_n(\mathbb{R})$ is the set of invertible upper triangular matrix of order n.

2.2 Compute of ideal context

In this section we attempt to solve the equation 3 which leads to the ideal context we look for.

If B is a solution of equation 3, then for all $1 \leq i \leq j \leq n$ we have

$$\frac{\partial(\frac{(\sum_{d \in R} d)^T.B^T.q}{(\sum_{d \in S} d)^T.B^T.q})}{\partial b_{ij}} = 0 \quad (4)$$

and so

$$\frac{\partial((\sum_{d \in R} d)^T.B^T.q)}{\partial b_{ij}}.(\sum_{d \in S} d)^T.B^T.q - \frac{\partial((\sum_{d \in S} d)^T.B^T.q)}{\partial b_{ij}}.(\sum_{d \in R} d)^T.B^T.q = 0 \quad (5)$$

that is

$$(\sum_{d \in R} d)^T.(\frac{\partial B}{\partial b_{ij}})^T.q.(\sum_{d \in S} d)^T.B^T.q - (\sum_{d \in S} d)^T.(\frac{\partial B}{\partial b_{ij}})^T.q.(\sum_{d \in R} d)^T.B^T.q = 0 \quad (6)$$

where

$$B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ 0 & b_{22} & \dots & b_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & b_{nn} \end{pmatrix}$$

The matrix $\frac{\partial B}{\partial b_{ij}}$ is:

$$\frac{\partial B}{\partial b_{ij}} = \begin{array}{cc} & \begin{array}{ccccccc} & & & j & & & \end{array} \\ \begin{array}{c} \\ \\ i \\ \\ \\ \\ \\ \end{array} & \begin{pmatrix} 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \end{pmatrix} \end{array}$$

Table 1. The ideal context maximizes the inner product between query vector and relevant document vectors.

documents	old inner product	new inner product
d_1	7	7.85
d_2	9	11.99

Table 2. The ideal context minimizes the inner product between query vector and irrelevant document vectors.

documents	old inner product	new inner product
d_3	2	0.98
d_4	1	0.19
d_5	2	1.22
d_6	1	0.22
d_7	2	1.2

2.3 Ideal query

Let C be the ideal context and q be the initial query.

$q' = C^T.q$ is called the *ideal query*.

3 An illustrative example

We consider the Topic 351 of the test collection TREC-7:

<top>

<num> Number: 351

<title> Falkland petroleum exploration

<desc> Description: What information is available on petroleum exploration in the South Atlantic near the Falkland Islands?

<narr> Narrative: Any document discussing petroleum exploration in the South Atlantic near the Falkland Islands is considered relevant. Documents discussing petroleum exploration in continental South America are not relevant.

</top>

Let $d_1 = (7,7,5,4,4,3)^T$ and $d_2 = (9,2,1,4,6,1)^T$ be two relevant documents.

Let $d_3 = (2,0,1,0,2,2)^T$, $d_4 = (1,0,0,1,1,2)^T$, $d_5 = (2,0,3,0,0,1)^T$, $d_6 = (1,0,3,0,0,1)^T$ and $d_7 = (2,0,0,0,0,1)^T$ five irrelevant documents.

These seven documents are described by six descriptors: falkland, british, company, island, air, water.

The initial query is: $q = (1,0,0,0,0,0)^T$.

By applying Equation 6 we obtain the following matrix which represents the ideal context:

$$B = \begin{pmatrix} 1 & 0 & 0.01 & 0.51 & 0.29 & -0.8 \\ 0 & 0.94 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0.8 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.75 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

This context maximizes the inner product between query vector and relevant document vectors and minimizes the inner product between query vector and irrelevant document vectors (Table 1 and Table 2).

Table 3. comparison between the ICM model, the BM model and the RM model.

	BM	RM	ICM	ΔBM	ΔRM
R-Precision	0.246	0.295	0.312	26.8 %	5.58%

Table 4. Comparison between the ICM model and the IRiX model.

	IRiX	ICM	Δ
P@5	0.332	0.4	20.48%
P@10	0.308	0.43	39.61%
P@15	0.276	0.41	48.55%

4 Experiments

In this section we give the different experiments and results obtained to evaluate our approach.

4.1 Environnement

The test collection TREC-7[1] is used in this study. The initial ranking of documents is weighted by the $BM25$ formula proposed in [8].

Let q be a query, d be a document and C be an ideal context matrix. The retrieved documents are ranked by the inner product:

$$(C.d)^T.q = d^T.C^T.q \tag{7}$$

4.2 Results

The experiments and the evaluations are articulated around the following two axes. First, the comparison between the Ideal Context Model (ICM) (equation 7), the Baseline Model (BM) [8] and Rocchio Model (RM) [5] using the R-Precision. Second, the comparison between the ICM model and the IRiX model [2] using P@5, P@10 and P@15. We show that our model improves results of [2] (Table 4).

5 Conclusion

This paper completes the papers [1, 2].

In this paper we define and compute an ideal context. This context guarantees an ideal representation of documents, that is the relevant documents are gathered and the irrelevant ones are kept away from the relevant ones. Different experiments and results show that our approach provides better results than the other ones.

In a forthcoming paper, similarity between context, in a vector space model, is studied then we hope compute an ideal context using a new similarity score.

[1] We test our algorithm using collection TREC-7 to compare our results with the previous studies [2].

References

[1] M. Melucci. Context modeling and discovery using vector space bases. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM), Bremen, Germany. ACM Press, 808-815, 2005.

[2] M. Melucci. A basis for information retrieval in context. ACM Trans. Inf. Syst., 26(3), 1-41, 2008.

[3] C.J. van Rijsbergen. The Geometry of Information Retrieval. Cambridge University Press UK, 2004.

[4] S. Robertson and K. Sparck-Jones. Relevance weighting of search terms. Journal of the American Society of Information Science, 129-146, May-June 1976.

[5] J. Rocchio. Relevance feedback in information retrieval. The SMART retrieval system-experiments in automatic document processing, 313-323, 1971.

[6] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. Communications of the ACM, 18(11), 613-620, November 1975.

[7] G. Salton. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison Wesley, 1989.

[8] Stephen, E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, Marianna Lau: Okapi at TREC. TREC 21-30, 1992.