Natural Language Engineering **18** (2): 155–185. © Cambridge University Press 2012
doi:10.1017/S1351324912000010

Document ranking refinement using a Markov random field model[*]

ESAÚ VILLATORO, ANTONIO JUÁREZ, MANUEL MONTES, LUIS VILLASEÑOR and L. ENRIQUE SUCAR

Department of Computer Science, National Institute of Astrophysics, Optics and Electronics, Luis Enrique Erro 1 Tonantzintla, Puebla, CP 72840, México
e-mail: {*villatoroe,antjug,mmontesg,villasen,esucar*}*@inaoep.mx*

(*Received 21 March 2011; accepted 28 December 2011*)

Abstract

This paper introduces a novel ranking refinement approach based on relevance feedback for the task of document retrieval. We focus on the problem of ranking refinement since recent evaluation results from Information Retrieval (IR) systems indicate that current methods are effective retrieving most of the relevant documents for different sets of queries, but they have severe difficulties to generate a pertinent ranking of them. Motivated by these results, we propose a novel method to re-rank the list of documents returned by an IR system. The proposed method is based on a Markov Random Field (MRF) model that classifies the retrieved documents as relevant or irrelevant. The proposed MRF combines: (i) information provided by the base IR system, (ii) similarities among documents in the retrieved list, and (iii) relevance feedback information. Thus, the problem of ranking refinement is reduced to that of minimising an energy function that represents a trade-off between document relevance and inter-document similarity. Experiments were conducted using resources from four different tasks of the Cross Language Evaluation Forum (CLEF) forum as well as from one task of the Text Retrieval Conference (TREC) forum. The obtained results show the feasibility of the method for re-ranking documents in IR and also depict an improvement in mean average precision compared to a state of the art retrieval machine.

1 Introduction

Information Retrieval (IR) deals with the representation, storage, organisation and access to information items[1] (Baeza-Yates and Ribeiro-Neto 1999). Given some query, formulated in natural language by a user, the IR system is suppose to retrieve and sort according to their relevance degree documents satisfying user's information needs (Grossman and Frieder 2004).

[*] This work was done under partial support of CONACyT project grants: 61335, CB-2008-106013 and CB-2009-134186.

[1] Depending on the context, items may refer to text documents, images, audio or video sequences.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

The word *relevant* means that retrieved documents should be semantically related to the user information need. Hence, one main problem of IR is determining which documents are, and which are not relevant. In practice, this problem is usually regarded as a *ranking* problem, whose goal is to define an ordered list of documents such that documents similar to the query occur at the very first positions.

Over the past years, IR Models, such as: Boolean, Vectorial, Probabilistic and Language models have represented a document as a set of representative keywords (i.e. index terms) and defined a *ranking* function (or retrieval function) to associate a relevance degree for each document with its respective query (Baeza-Yates and Ribeiro-Neto 1999; Grossman and Frieder 2004). In general, these models have shown to be quite effective over several tasks in different evaluation forums (CLEF[2] and TREC[3]). Nevertheless, these retrieval systems still fail at retrieving most of the relevant documents to a given query in the first positions. The latter is due to the fact that modelling user intentions from queries is, in general, a highly subjective and difficult task, hence, post-processing and ranking refinement strategies have been adopted (Yang, Ji and Tang 2006; Sarkar and Moore 2009; Villatoro-Tello, Montes-y-Gómez and Villaseñor-Pineda 2009a, 2009b; Chávez, Sucar and Montes 2010; Zhou *et al.* 2010a, 2010b).

Post-retrieval techniques aim at refining retrieval results by feature re-weighting, query modification, document re-ranking and relevance feedback. The common idea is to interact with the user in order to learn or to improve a model of the underlying user's information need. Acceptable results have been obtained with such methods, however, they still have several limitations, including: (i) the need of extensive user interaction[4]; (ii) multiple execution of retrieval models; (iii) the on-line construction of classification methods; (iv) the lack of contextual information in the post-retrieval processing, which may be helpful for better modelling users' information needs; and (v) the computational cost that involves processing the entire collection of documents each feedback iteration.

1.1 Our approach

This paper introduces an alternative post-retrieval technique that aims at improving the results provided by a document retrieval system and that overcomes some of the limitations of current post-retrieval methods. In particular, we face the problem of re-ranking[5] a list of documents retrieved by some IR system. This problem is motivated by the availability of retrieval systems that present high-recall and low-precision performance, which evidences that the corresponding retrieval system is in fact able to retrieve many relevant documents but has severe difficulties to generate a

[2] Cross Language Evaluation Forum (http://www.clef-campaign.org/).
[3] Text Retrieval Conference (http://trec.nist.gov/).
[4] It is worth mentioning that if available, user interaction should be included in post-retrieval techniques as it is evident that information provided by the user is much more reliable than that obtained by a fully automatic process. Hence, the goal of post-processing techniques should be minimising users' interaction instead of completely eliminate it from the process.
[5] Also known as the problem of *Ranking Refinement*.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

pertinent ranking of them. Hence, given a list of ranked documents, the problem we approach consists of moving relevant documents to the first positions and displacing irrelevant ones to the final positions in the list.

We propose a solution to the ranking refinement problem based on a Markov Random Field (MRF) (Kemeny, Snell and Kanpp 1976; Pearl 1988; Lauritzen 1996; Winkler 2006) that aims at classifying the ranked documents as relevant or irrelevant. Each document in the retrieved list is associated to a binary random variable in the MRF (i.e. a node), the value of each random variable indicates whether a document is considered relevant (when its value is 1) or not (when its value is 0). The MRF considers several aspects: (1) the information provided by the base IR system, (2) similarities among retrieved documents in the list and (3) information obtained through a relevance feedback process. Accordingly, we reduce the problem of ranking refinement to that of minimising an energy function that represents a trade-off between document relevance and inter-document similarity. The information provided by the IR system is the base of our method, which is further enriched with contextual and relevance feedback information.

Our motivation for considering context information is that relevant documents to a query will be similar to each other and to its respective query, to some extent; whereas irrelevant documents will be different among them and not as similar to the query as the relevant documents[6]. Relevance feedback information has two main purposes: (i) to work as a seed generation mechanism for propagating the relevancy/irrelevancy status of nodes (documents) in the MRF, and (ii) to denote the users' search intention by working as *example texts*.

At this point, it is important to mention that, traditionally a relevance feedback process takes as input a set of n documents (tentatively relevant) and generates as output a set of k isolated terms (tentatively relevant to the query) which are further employed for a query expansion (QE) process. For our purposes, we will employ the complete documents (called *example texts*), since by doing this, we have showed (Villatoro-Tello *et al.* 2009a, 2009b) that it is possible to make a more accurate approximation of the users' search intention (i.e. to become into a more explicit representation the implicit information contained in the query).

The proposed MRF does not require of multiple executions of IR models, nor training classification methods, and it can work without user intervention; therefore, our MRF overcomes the main limitations of current post-processing techniques; see Section 2.

We report experimental results in four CLEF collections (Geo CLEF 2008, Ad Hoc CLEF 2005, Robust CLEF 2008, Image CLEF 2008) and one TREC collection (TREC-9 SDR) that show the pertinence of our method. Experimental results show that our method is able to improve the ranking provided by the IR system. Statistical tests were also performed to prove that obtained results were in fact significant, motivating us for further research.

[6] Keep in mind that irrelevant documents will be similar to the query in some degree since such documents were obtained by an IR system through that query in the first place.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

1.2 Structure of the paper

The rest of this papers is organised as follows. The next section describes some of the recent work on ranking refinement in document retrieval. Section 3 presents some background concepts that will help the reader for a better understanding of the proposed method. Section 4 introduces the proposed MRF for ranking refinement in document retrieval. Section 5 describes the main characteristics of employed collections as well as our base IR system. In Section 6, experimental results are discussed, and in Section 7, some set of additional experiments are presented with the aim of comparing our proposed method against one of the most common re-ranking strategies, QE. Finally, Section 8 depicts the main conclusions derived from this work and outlines future work directions.

2 Related work

Document re-ranking or ranking refinement in IR has been a widely research topic during the last fifthteen years. There are two main approaches for this task: (i) indirect re-ranking via some QE strategy, and (ii) direct re-ranking on initial retrieved documents (Yang *et al.* 2006). Normally, QE strategies assume that top ranked documents are more likely to be relevant, the terms contained within these documents can be used to augment the original query and then a better ranking can be expected via a second retrieval process. In contrast, direct re-ranking strategies try to improve the ranking of the initial set of retrieved documents by directly adjusting their positions without the need of performing a second retrieval process, normally, this type of strategy use the information contained within the retrieved documents (e.g. inter-document similarities) to generate a better ranking of them. The generated output (i.e. a list of ranked documents) by any of this two strategies would be of obvious benefit to users, for example, direct ranking refinement can be used to improve automatic QE since a better ranking in the top retrieved documents can be expected.

Our work, as we mention in Section 1.1, focuses on direct document ranking refinement. Depending on the employed information source, direct document ranking refinement can be classified into the following categories.

Based on inter-document relationships. As an example of this type of approaches, (Balinski and Danilowicz 2005) generate a re-ranked list of documents by using documents distances to modify their relevance weight, while Lee, Park and Choi (2001); Yang *et al.* (2006); Bendersky and Kurland (2008) and Sarkar and Moore (2009) proposed a ranking refinement approach based on document clustering, where the key idea is that nodes relatively closer (in graph topology) to some set of presumable relevant nodes are more likely to be relevant.

Based on external resources. One of the main disadvantage of this approach is the high dependence on the availability of various external resources, such as manually built thesaurus (Qu, Xu and Wang 2000) where the purpose of the thesaurus is to enrich queries and then a re-ranking process is performed by

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

means of computing similarities of each retrieved document to its respective enhanced query. In (Zhou *et al.* 2010a, 2010b) they try to directly model the internal structure of topics by means of a weighted vector of knowledge-based concepts derived from external resources such as Wikipedia articles, hence, if a group of documents deal with the same enhanced topic, these documents will get allocated similar ranking as they are more likely to be relevant to the query. Other example of this type of approaches are (Bear *et al.* 1997), which represents queries through manually crafted grammars; and (Kamps 2004; Yang *et al.* 2004) that represent documents employing a set of key-phrases or key-terms, which are extracted from a controlled vocabulary, and try to re-rank the set of initially retrieved documents employing its respective type of representations by computing similarities among queries and documents.

Based on common characteristics found both in documents and queries. The main characteristic of this type of methods is the use of specific information extracted from documents or queries. For example, (Luk and Wong 2004) used information in the document title; (Crouch *et al.* 2002) used stemmed words in the initial retrieval and augmented un-stemmed words in queries in document re-ranking; (Xu and Croft 1996, 2000) made use of global and local information to re-rank the documents via local context analysis; (Mitra, Singhal and Buckley 1998) used maximal marginal relevance to adjust the contribution of relevant terms; (Yang and Ji 2005a, 2005b) used query terms, which occur in both queries and top retrieved documents, to re-rank documents.

Based on graph structure. This type of methods represent retrieved documents in a graph topology in order to explore their intrinsic structure and then to produce a re-rank of them. (Kurland and Lee 2005) performed re-ranking based on measures of centrality in the graph formed by the generation of links induced by language model scores, through a weighted version of the PageRank algorithm and a HITS-style cluster based approach. In (Zhang *et al.* 2005), a similar method is employed to improve web search based on a linear combination of results from text search and authority ranking. The graph, which it is called "affinity graph", shares strong similarities with (Kurland and Lee 2005) where the links are induced by a modified version of cosine similarity using the vector space model. (Diaz 2005) used score regularisation to adjust document retrieval ranking from an initial retrieval by a semi-supervised learning method. (Deng, Lyu and King 2009) further developed this method by building a latent space graph based on content and explicit link information, similar to this method, except that they model the explicit information in a direct form as proposed by (Zhou *et al.* 2010a, 2010b). Finally, the method proposed by (Yang *et al.* 2006) adds to the graph structure of documents, the use of learning algorithms to perform a re-ranking considering the intrinsic information among top retrieved documents. (Yang *et al.* 2006) employs a label propagation-based learning algorithm to integrate pseudo labeled data with unlabeled data. Their proposed algorithm first represents labeled and unlabeled examples as vertices in a connected graph, then propagates the label

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

information from any vertex through weighted edges and finally infers the labels of unlabeled examples until the propagation process converges.

Whereas most of the above cited works have reported acceptable performance, they still have several limitations. Some of them require the execution of a retrieval model multiple times, the on-line constructions of classification methods or the need of external resources for acquiring terms semantically related to queries and/or documents; all of these formulations can be computationally expensive. Additionally, most of the described methods do not take into account all of the available information (e.g. initial ranking and contextual information) for refining the retrieval results of the base IR system, which can be helpful for improving the effectiveness of post-retrieval techniques.

In this work, we propose a direct ranking refinement strategy for improving the output generated by a document retrieval system. Our approach does not require the multiple execution of a retrieval model, the construction of a classifier nor the use of external resources such as thesaurus. For this purpose, we propose a variant of relevance feedback that aims at overcoming the limitations of current methods by considering full documents (called *example texts*) instead of a set of k isolated terms. By providing full *example texts* as feedback elements, it is possible to generate a more explicit approximation of the implicit information contained in a given query. The proposed method do not need to process the entire collection of documents in each feedback iteration, but it focuses on the top n retrieved documents. It is also important to mention that our method requires minimal user intervention and it can even work without the need of a user at all (e.g. under a blind relevance feedback formulation). Additionally, the proposed model takes advantage of all of the information available during a retrieval session, namely: initial ranking as provided by the base IR system, inter-document similarity among retrieved elements (i.e. context) and relevance feedback information. A notable benefit of our approach is that it is not restricted to a particular retrieval system or system architecture. Thus, our method offers advantages in terms of generality, as it can be used with any retrieval system, robustness, as it does not require an user-in-the-loop, as well as efficiency and effectiveness.

3 Background information

In this section, we provide some background concepts that will be helpful for understanding the rest of this paper.

3.1 Query expansion via relevance feedback

A QE via relevance feedback process is a controlled technique which main goal is to reformulate a query. In other words, a relevance feedback strategy is normally a previous step for a QE process. The basic idea is to select a set of k words which are related to a set of documents that have been previously retrieved and tagged as relevant by some user. Further, these words are added to the original query (Salton and Buckley 1990). In order to apply a relevance feedback process, it is necessary to

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

perform a first search (i.e. a first retrieval process) which generates an ordered list of documents. Afterwards, the user selects from the first positioned documents those that he considers as relevant (i.e. the user establishes the documents' relevance). This relevance judgements that the user just gave to the documents are employed to compute a new set of values that indicate in a more accurate form the impact of each word in the original query[7].

Commonly, the k most frequent words within the documents tagged as relevant are considered for its addition to the original query. One of the main advantages of the QE via relevance feedback process is that usually the IR systems are able to obtain better *precision* and *recall* levels as long as the added words are *good* terms, i.e. relevant words to the user's information need. Among the disadvantages is the computational cost implied, since it is necessary to perform a second retrieval process. Besides this, relevance feedback strategies have shown to be sensitive to the quality of the added words, since adding an *irrelevant word* could be very harmful for the IR system.

3.2 Markov random fields

MRF are a type of undirected probabilistic graphical models that aim at modeling dependencies among variables of the problem in turn (Kemeny *et al.* 1976; Pearl 1988; Lauritzen 1996; Winkler 2006). MRFs have a long history within image processing and computer vision (Li 1994, 2001). They were first proposed for denoising digital images (Kemeny *et al.* 1976; Pearl 1988; Lauritzen 1996; Winkler 2006) and since then a large number of applications and extensions have been proposed. Classical applications include image segmentation (Held *et al.* 1997) and image filtering (Geman and Geman 1984); although, recently they have been successfully applied for image annotation (Carbonetto, De Freitas, and Barnard 2004), region labeling (Escalante, Montes and Sucar 2007; Hernández and Sucar 2007) and IR (Metzler and Croft 2005, 2007; Lease 2009) with great success.

MRF modeling has appealing features for problems that involve the optimisation of a configuration of variables that have inter-dependencies among them. Accordingly, MRFs allow the incorporation of contextual information in a principled way. MRFs rely on a strict probabilistic modeling, yet they allow the incorporation of prior knowledge by means of potential functions. For those reasons, in this paper, we adopted an MRF model for refining the initial ranking of a set of documents retrieved by some IR system. The rest of this sections summarises the formalism of MRFs.

An MRF is a set of random variables $F = \{f_1, \ldots, f_N\}$ indexed by sites or nodes where the following conditions hold:

$$P(f_i) \geq 0, \forall f_i \in F \tag{1}$$

$$P(f_i | f_{S-\{i\}}) = P(f_i | \mathcal{N}(f_i)) \tag{2}$$

[7] When the relevant documents are identified by some automatic process, it is assumed that documents placed at the top positions of the list are in fact relevant, and the new set of words that will be added to the query are automatically selected; this type of feedback is known as *blind relevance feedback*.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

where $\mathcal{N}(f_i)$ is the set of neighbours of f_i according to the neighbouring system $\mathcal{N}$. Formula 1 is the so called positivity condition and avoids negative probability values, whereas expression 2 states that the value of a random variable depends only on the set of neighbours of that variable.

It has been shown that an MRF follows a Gibbs distribution (Geman and Geman 1984), where a Gibbs distribution of the possible configurations of F with respect to $\mathcal{N}$ has the following form:

$$P(F) = Z^{-1} \times e^{-\frac{1}{T}E(F)} \tag{3}$$

where Z is a normalisation constant and the T is the so called temperature parameter (a common choice is $T = 1$) and $E(F)$ is an energy function of the following form:

$$E(F) = \sum_{c \in C} V_c(f) = \sum_{\{i\} \in C_1} V_1(f_i) + \sum_{\{i,j\} \in C_2} V_2(f_i, f_j) + \ldots \tag{4}$$

where "..." denotes possible potentials V_c defined over higher order neighbourhoods $C_3, C_4 \ldots, C_K$; each C_i defines a neighbourhood system of order i between the nodes of the MRF. Often, the set F is considered the union of two subsets of random variables $X \cup Y$; where X is the set of observed variables and Y is the set of output variables, which state that we would like to predict. Potentials V_c are problem dependent and commonly learned from data.

One of the main problems in MRFs is that of selecting the most probable configuration of F (i.e. an assignment of values to each variable f_i of the field). Such configuration is determined by the configuration of F that minimises expression 4, for which a diversity of optimisation techniques have been adopted (Kemeny *et al.* 1976; Pearl 1988; Lauritzen 1996; Winkler 2006).

4 Proposed Markov random field for ranking refinement

As we previously mention, we focus on the problem of ranking refinement based on the fact that current IR methods are effective retrieving most of the relevant documents for different sets of queries, but they have severe difficulties generating a pertinent ranking of them. Our main hypothesis establishes that for any set of retrieved documents, obtained as a result from a common query, there must exist some degree of homogeneity among the relevant documents, whereas irrelevant documents tend to be relatively heterogeneous. Hence, our idea is to incorporate inter-document similarities, position (i.e. *rank*) of the documents in the original retrieved list, and similarities against a set of *example texts* in a MRF to generate the final re-ranked list of documents.

The MRF we propose takes as input a list of D-ranked documents, which are provided by an IR system, and attempts to re-rank the documents in the list in such a way that relevant documents are placed before irrelevant ones. The proposed method can be added as a post-processing stage for any IR system, as it does not rely on information from a particular system. The rest of this section describes the MRF we propose.

We consider an MRF in which each variable $F = \{f_1 \ldots, f_N\}$ corresponds to a document d_i in the list D returned by the base IR system; each f_i is a binary

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

random variable such that when $f_i = 1$ the *ith*document is considered to be relevant to the search intention and when $f_i = 0$ the corresponding document is considered to be irrelevant. The main task of the MRF is to divide the documents in the list into relevant and irrelevant ones by varying the values of $F = \{f_1, \ldots, f_N\}$. Based on the final configuration of the MRF, we generate a new list of documents by placing in the first positions those documents i for which $f_i = 1$, followed by the rest of the documents (i.e. documents with $f_i = 0$). We define an energy function for the MRF that attempts to model the relevancy status of documents in terms of: (i) the rank information provided by the base IR system, (ii) the similarity against a set of *example texts* obtained through some relevance feedback strategy; and (iii) similarities among documents in the retrieved list.

4.1 Energy function

The energy function has two main parts, the *interaction potential* and the *observation potential*; it has the following form:

$$E(F) = \lambda \left(\sum_{f_i \in F} V_c(f_i, N_i) \right) + (1 - \lambda) \left(\sum_{f_i \in F} V_a(f_i) \right) \tag{5}$$

where N_i is the set of neighbours of node f_i, V_c is the interaction potential and accounts for information of the association between surrounding documents, whereas V_a is the observation potential and it accounts for information that is associated to a single document (see Figure 1). λ is a scalar that weights the importance of both V_a and V_c.

The energy function of the MRF specifies how likely a configuration of the MRF (i.e. an assignment of values to each node f_i) is the best ranking for documents in the list. We define an energy function that incorporates inter-document similarity, also the similarity against a set of *example texts* which are obtained through a relevance feedback process and rank information provided by the base retrieval system. The underlying idea is that a combination of these information sources is beneficial for characterising *good ranks*.

As final remark, observe that both elements of the energy function (Expression 5) are quadratic; i.e., $O(V_c) = O(n^2)$ and $O(V_a) = O(n^2)$ being n the number of documents contained in the retrieved list. Hence, $O(E) = O(n^2) + O(n^2) = O(max(n^2, n^2)) = O(n^2)$ and since λ is a constant, it is a depreciable element from the complexity calculation.

Next, we provide a general description on how the energy function is computed and in Section 4.2, we provide details on how does the dissimilarity between documents is estimated.

4.1.1 Interaction potential

Intuitively, V_c asses how much support provide the neighbouring same-valued variables to f_i so it keep its current value, and also how much support give oppose-value variables to f_i so it changes to the contrary value. The interaction

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

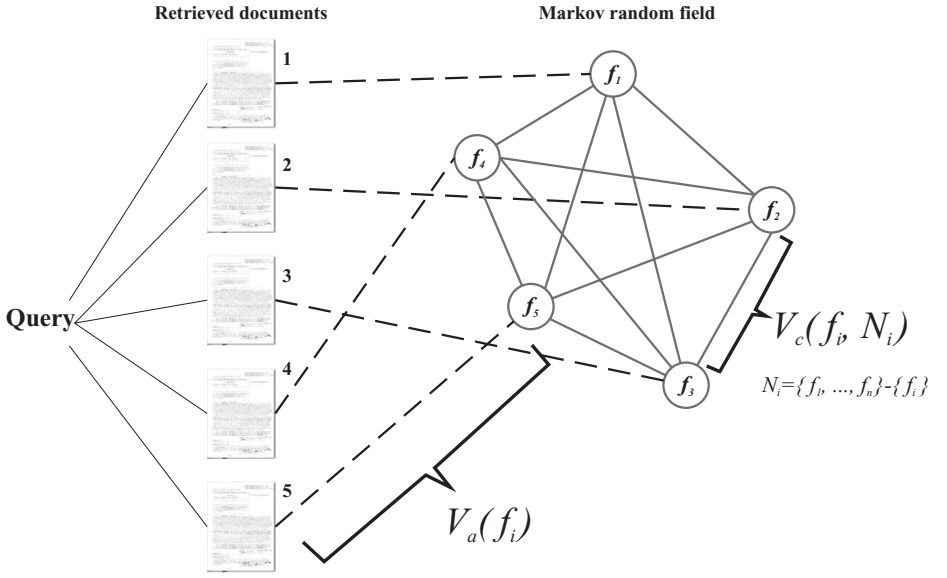


Fig. 1. Diagram of the proposed MRF for document re-ranking. Each document in the original retrieved list is associated to a node in the MRF. Single node information is incorporated through the $V_a(f_i)$ potential (dashed lines), while contextual information is introduced via the $V_c(f_i, N_i)$ potential (solid lines).

potential V_c is defined as:

$$V_c(f_i, N_i) = \begin{cases} g\left(f_i, N_i^I\right) + \left(1 - g\left(f_i, N_i^R\right)\right), & \text{if } f_i = 0 \\ g\left(f_i, N_i^R\right) + \left(1 - g\left(f_i, N_i^I\right)\right), & \text{if } f_i = 1 \end{cases} \tag{6}$$

where $g(f_i, N_i^I)$ is the average dissimilarity between the document associated with the node f_i and its neighbours with irrelevant value N_i^I. Conversely, $g(f_i, N_i^R)$ represents the average dissimilarity between the associated document to f_i and its neighbours with relevant value N_i^R. Thus, we divide the neighbours of node f_i into two subsets: neighbours with relevant value (N_i^R), and neighbours with irrelevant value (N_i^I)[8].

Accordingly, we define the dissimilarity measure $g(f_i, N_i^X)$ for a subset of relevant (R) or irrelevant (I) nodes N_i^X (being X a place-holder for R or I) as:

$$g(f_i, N_i^X) = \frac{\sum_j^{|N_i^X|} dist(d_i, d_j)}{|N_i^X|} \tag{7}$$

where $dist(d_i, d_j)$ represents a distance function between d_i and d_j, which is computed by means of comparing the textual features of the documents (see Section 4.2). Remember that documents d_i and d_j are associated to nodes f_i and f_j respectively, accordingly, the dissimilarity measure (g) between a node f_i and its neighbours N_i^X

[8] Notice that we assume that each node in the MRF is connected to each other, this is, a fully connected graph, and therefore $N_i = F - \{f_i\}$. However, this fact is not a problem since our approach considers using a reduced number of nodes.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

is defined as the normalised sum of the distance between the associated document d_i and all its surrounding documents.

It is important to mention that the idea of rewarding irrelevant documents comes from the fact that retrieved documents were obtained by means of the same query. Although we assume that irrelevant documents will be relatively heterogeneous, these documents will contain small similarities among them because of this fact, i.e. they may have a minor degree of homogeneity. Consequently, by rewarding these *small* similarities our MRF model tries to distinguish relevant from irrelevant documents in the list.

4.1.2 Example texts and the virtual document

As we mentioned in Section 1.1, the information obtained by the relevance feedback process is a set of *example texts* instead of a set of k isolated terms. In previous work (Villatoro-Tello *et al.* 2009a, 2009b), we have shown that by using *example texts* it is possible to generate a more explicit approximation of the users' information need.

In order to be able to process the information contained in the *example texts*, we define a virtual document v as the result of concatenating $|S|$ *example texts* which were previously obtained by a relevance feedback strategy. Formally, having a set $D = \{d_1, d_2, \ldots, d_{|D|}\}$ which is an ordered list of $|D|$ retrieved documents by some IR system, we will select from D a total of $|S|$ example texts by applying some relevance feedback strategy, generating the set of example texts $S = \{s_1, s_2, \ldots, s_{|S|}\}$.

Then, to define our virtual document v, we first concatenate all the *example documents*, and then we generate its representation using the traditional *tf-idf* weighting scheme (see Section 4.2).

4.1.3 Observation potential

The observation potential V_a is based on the assumption that relevant documents are similar to the set of *example texts* and at the same time it is likely that they appear in the top positions; alternatively, it assumes that irrelevant documents are less similar to the set of *example texts* and it is very probably that they appear at the bottom positions of the list. The observation potential V_a is defined as follows:

$$V_a(f_i) = \begin{cases} (1 - dist(d_i, v)) \times \delta(1/rank(d_i)), & if\ f_i = 0 \\ dist(d_i, v) \times \delta(rank(d_i)), & if\ f_i = 1 \end{cases} \tag{8}$$

V_a captures the affinity between the document d_i associated to node f_i and the *virtual document* v, as measured by a distance term and by using the rank information of the original list. In the one hand, $dist(d_i, v)$ indicates how distant is the document d_i to the virtual document v. On the other hand, δ is a function that transforms the position ($rank(d_i)$) of the document d_i in the original list into a real value (Chávez *et al.* 2010).

The latter factor incorporates information from the initial retrieval system; however, notice that we only use the position of the documents in the list, which is

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

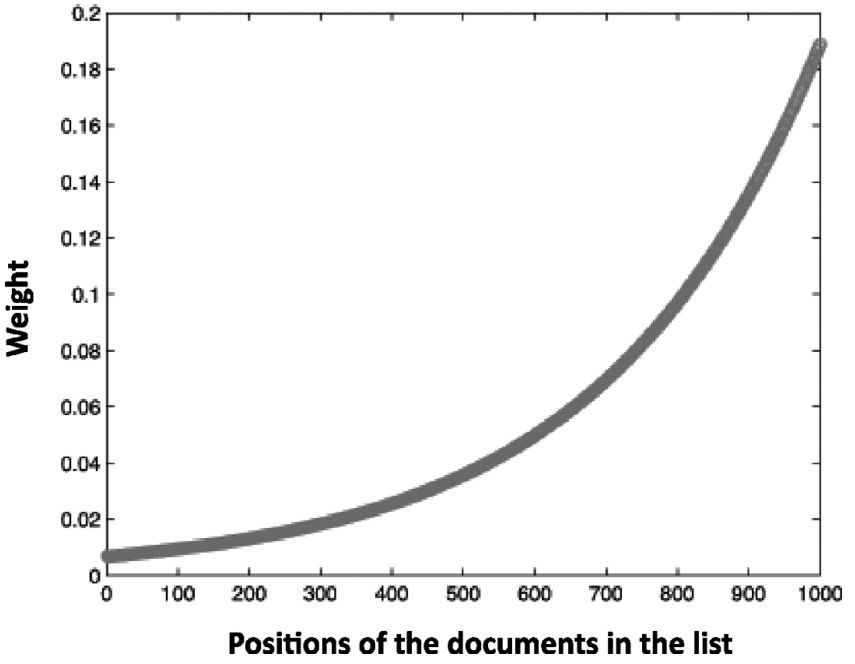


Fig. 2. Graphical behaviour of the transformation δ when $c_1 = 300$ and $c_2 = 5$. Observe that $\delta(x)$ will take a value proportional to the position of the document in the list.

an independent value of the retrieval system that was employed to obtain the initial list.

The transformation δ is defined as follows:

$$\delta(x) = \frac{\exp\left(\frac{x}{c_1}\right)}{\exp(c_2)} \tag{9}$$

when $f_i = 1$, $\delta(x)$ takes values proportional to the position of the document in the list; when $f_i = 0$, $\delta(x)$ takes values proportional to the inverse of the position in the list. The position (inverse of the position) of documents is weighted exponentially because we want the position (inverse of the position) to have more influence for the top-ranked (bottom-ranked) documents. This function was inspired on the distance based probability model due to (Mallows 1975): $\beta(x) = \frac{\exp(\theta; g(x,\Pi))}{\phi}(\theta)$.

In the one hand, constant c_1 in expression 9 represents the point from which we will start assigning a greater penalisation value. For our experiments c_1 was set to 300. On the other hand, constant c_2 reflex the scale of the penalisation values, i.e. to greater values of c_2 lower penalisation scores, and vice-versa. For performed experiments, c_2 was set to 5. Figure 2 resumes the behaviour of δ for the selected values.

4.1.4 Defining the MRF's initial configuration

As we previously mention, we employed the relevance feedback information for two main purposes: (i) to work as a seed generation mechanism for propagating the

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

relevancy/irrelevancy status of nodes (documents) in the MRF, and (ii) to denote the users' search intention by working as *example texts*. The latter has been already explained in Section 4.1.2, in this section, we will explain the first one.

We use relevance feedback information as a seed for building the initial configuration of the MRF; that is, we set $f_i = 1$ for relevance feedback documents (i.e. the *example texts*) and we set $f_j = 0$ for the rest. In this way, the MRF starts the energy minimisation process knowing what documents are potentially relevant to the users' information need. Thus, the inference process consists of identifying further relevant documents in the list by propagating through the MRF the relevance feedback information. Since, we assume that *example texts* are indeed relevant to the users' information need, they are considered to be relevant during the whole energy minimisation process (i.e. the corresponding nodes are never set to 0).

4.1.5 Inference in the MRF

As stated before, the configuration that minimises expression 5 is used for generating the new rank of documents. Such configuration is obtained via stochastic simulation using the iterated conditioned modes (ICM) algorithm (Besag 1986). In the future, we will explore the use of other strategies.

The ICM uses a deterministic "greedy" strategy to find a local minimum. It starts with an estimate of the labelling, which is given by the *example texts* (See Section 4.1.4), and then for each node, the label which gives the largest decrease of the energy function is chosen. This process is repeated until convergence, which is guaranteed to occur, and in practice is very fast (Besag 1986).

It is important to mention that for this work, our purpose is that it was not to compare several optimisation process neither in terms of solution quality or running time efficiency. Rather than that, our goal was to provide some evidence of the pertinence of using MRF's for a ranking refinement process.

4.2 Estimation of similarities

In Section 4.1, we described the form of the energy function we use. In this section, we describe how the dissimilarity between documents is estimated. The proposed dissimilarity measure is basically a normalised number representing the overlap of words between documents that are compared.

4.2.1 Textual features

As textual features, we use a bag of words representation, where each document is represented by a weighted vector that indicates the presence and importance of words from the collection vocabulary in the document. This type of representation is known as the vectorial space model representation (VSM) (Salton, Yang and Wong 1975), which basic idea establishes that the main meaning of a document is given by the words contained in it. This type of representation proposes to transform documents to a vectorial representation by employing the words contained in it.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

In the VSM model, each document d_i is represented by a vector of length equal to the vocabulary size $|V|$. The collection vocabulary is the set of all different terms (i.e. words) that appear in the collection of documents. Each component k from the vector indicates the contribution of term k within the document represented by that vector. The set of vectors representing documents within the collection generate a vectorial space where documents can be compared through its vectorial representation. Then, the generated vectorial space is represented by the matrix (M^{TD})[9], also known as term-document (TD) matrix of size $M \times N$, where M is equal to the size of the vocabulary (i.e. $M = |V|$), and N is the number of documents within the collection. Each entry M_{ik}^{TD} represents the contribution of term t_k within document d_i.

Several weighting schemes have been proposed, however the most employed is the *term-frequency inverse-document-frequency* (*tf-idf*) scheme. This strategy assigns to each value of M_{ik}^{TD} the value determined by the following expression:

$$M_{ik}^{TD} = tf_{ik} \times \log\left(\frac{|D|}{df_k}\right) \tag{10}$$

where tf_{ik} represents the number of times term k appears in document d_i. $|D|$ is the total number of documents and df_k is the number of documents that contain term k.

4.2.2 Similarity of textual features

In order to compute the similarity between documents, we employed the following functions:

- The dissimilarity between two documents d_i and d_j, which are associated to nodes f_i and f_j respectively is defined as follows:

$$dist(d_i, d_j) = 1 - sim(d_i, d_j) \tag{11}$$

  where $sim(d_i, d_j)$ represents a similarity function. For our experiments, we employed the well known *dice* coefficient, which is defined as follows:

$$sim(d_i, d_j) = \frac{2\sum_{k=1}^{|d_i|} d_{ik} d_{jk}}{\sum_{k=1}^{|d_i|}(d_{jk})^2 + \sum_{k=1}^{|d_i|}(d_{ik})^2} \tag{12}$$

  Hence, d_{ik} and d_{jk} represents the *tf-idf* value of the k^{th}-term of d_i and d_j respectively.
- The dissimilarity between a document d_i associated to node f_i and the virtual document v is defined in a similar form:

$$dist(d_i, v) = 1 - sim(d_i, v) \tag{13}$$

  where $sim(d_i, v)$ is also computed employing the *dice* coefficient.

[9] Following the notation presented in Section 4.1.2, we employ D to represent a set of documents and d_i indicating a particular document.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

5 Experimental setup

In this section, we describe the experimental setup that we employed to analyse the feasibility of the proposed method. A brief description of the base IR system used is given as well as its configuration. Besides this, we describe the test collections and the different sets of queries used to evaluate the proposed ranking refinement method.

5.1 Base IR system

As we have mentioned before, our ranking refinement strategy does not depend on any particular IR system. However, in order to perform our experiments we employed as base IR system the wellknown IR system LEMUR-INDRI. This system is part of the Lemur Project[10] started in 2000 by the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts, Amherst, and the Language Technologies Institute (LTI) at Carnegie Mellon University. Particularly, the LEMUR-INDRI toolkit is a search engine that provides state-of-the-art text search facilities, a rich structured query language for different text collections, and is considered as a robust system capable of producing comparable results to new IR schemes.

For all our experiments, the collections were indexed by this tool. For this purpose, collections were preprocessed by applying stop word elimination as well as a stemming process. For our experiments, we employed a list of 571 stop words available in the CLEF site[11]. Additionally, for the stemming process we employed the well known Porter algorithm (Porter 1997).

As baseline results, we considered the performance obtained under this configuration employing the LEMUR-INDRI search engine.

5.2 Collections

For our experiments, we employed four different collections, two of them are news documents, one is of annotated images (images with textual descriptions of their content), and finally one collection that consists of automatically generated transcriptions from news broadcasting. The first three collections are part of the CLEF evaluation forum and are employed for different tasks, the last one is part of the TREC forum. Following we give a brief description of these collections.

LA Times 94 (LA94) News from year 1994 written in American English, contains 113005 news documents in a total of 425Mb (Di Nunzio *et al.* 2008).

Glasgow Herald 95 (GH95) News from year 1995 written in British English, contains 56472 news documents in a total of 154Mb (Di Nunzio *et al.* 2008).

IAPR TC-12 Photographic collection of images from around the world, it includes photos of sports, persons, animals, cities, etc. It contains 20,000 images, each with it respective textual description (Grubinger 2007).

[10] http://www.lemurproject.org
[11] Cross Language Evaluation Forum (http://www.clef-campaign.org/).

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

Table 1. *Characteristics of the employed set of topics*

Task	Short name	Num. topics	Supported	Collection
Geo CLEF 2008	Geo08	25	24	LA94, GH95
Ad Hoc CLEF 2005	AdHoc05	50	50	LA94, GH95
Robust CLEF 2008	Robust08	160	153	LA94, GH95
Image CLEF 2008	Image08	39	39	IAPR TC-12
2000 TREC-9 SDR	SDR09	50	50	TREC-9 SDR

```
<top>
 <num>GC030</num>
 <EN-title>Car bombings near Madrid</EN-title>
 <EN-des>Documents about car bombings occurring near Madrid</EN-desc>
 <EN-narr>Relevant documents treat cases of car bombings occurring in the
capital of Spain
 and its outskirts</EN-narr>
</top>
```

Fig. 3. General structure of CLEF topics.

TREC-9 SDR A set of automatically generated audio transcriptions from several news broadcasting sites. Contains 21754 news transcriptions in a total of 91Mb (Garafolo, Auzanne and Voorhees 2000).

5.3 Topics

Figure 3 shows the general structure of CLEF topics. The main query or title is between labels <EN-title> and </EN-title>. Also a brief description (<EN-desc>, </EN-desc>) and a narrative (<EN-narr>, </EN-narr>) are given. These last two fields usually contain more information about the requirements of the original query. It is worth mentioning that TREC-9 SDR topics are formed by only one sentence or title, i.e. they do not have any description or narrative fields.

For our experiments, we worked with topics from four different tasks from the CLEF forum and one task from the TREC forum. It is important to mention that for all our experiments, we used the *title* and *description* fields only. Table 1 shows some information about the considered topics and the search collections employed by each task.

The *Supported* column in Table 1 indicates how many topics have at least one relevant document within their respective collection. On the other hand, the *short name* column shows the form in how we will refer from here to each task. Another important aspect of topics is the number of relevant elements that they have within the collection. Commonly, if a topic has many relevant elements, it is said that this topic is well supported by its respective collection. Table 2 shows some information about the number of relevant elements for each set of topics we employed. Generally, topics with many relevant elements have higher probabilities

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

Table 2. *Relevant elements contained in the search collections*

Task	Max.	Min.	Average	Standard Deviation
Geo08	109	1	31.12	31.69
AdHoc05	229	3	41.26	41.92
Robust08	229	1	28.28	34.04
Image08	184	18	61.56	33.73
SDR09	220	4	44.32	49.83

of satisfying its information need. Hence, topics with very few relevant elements are considered more difficult.

5.4 Evaluation

The evaluation of results was carried out using a measure that has demonstrated its pertinence to compare IR systems, namely, the Mean Average Precision (MAP). MAP is defined as follows:

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\left(\frac{\sum_{r=1}^{m} P_i(r) \times rel_i(r)}{n}\right)$$

Where $P_i(r)$ is the precision at the first r documents, $rel_i(r)$ is a binary function which indicates if document at position r is relevant or not for the query i; n is the total number of relevant documents for the query i, m is the number of relevant documents retrieved and Q is the set of all queries.

Intuitively, this measure indicates how well the system puts into the first positions relevant documents. It is worth pointing out that since our IR system was configured to retrieve 1,000 documents per query, MAP values are measured at 1,000 documents.

6 Experimental results

We conducted several experiments for evaluating the performance of the MRF model described in Section 4. The goal of the experiments were to: (i) evaluate the proposed MRF performance under different parameter settings; (ii) compare the performance of the MRF to that of the base IR system; and (iii) analyse the results of the proposed MRF when different relevance feedback strategies are considered. The rest of this section describes the obtained results and highlights our main findings.

Each run of our experiments proceeds as follows. First, we provide the original list as input to our MRF model. Next, through some relevance feedback strategy, a set of k *example documents* are selected from the list. These k *example documents* are used as the initial configuration for the method under evaluation, and for the construction of our virtual document v. The ICM algorithm is run up to 500 iterations trying to minimise the corresponding energy function. The obtained configuration of the method is employed to generate a new list of ranked documents, and finally this new generated list is evaluated.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

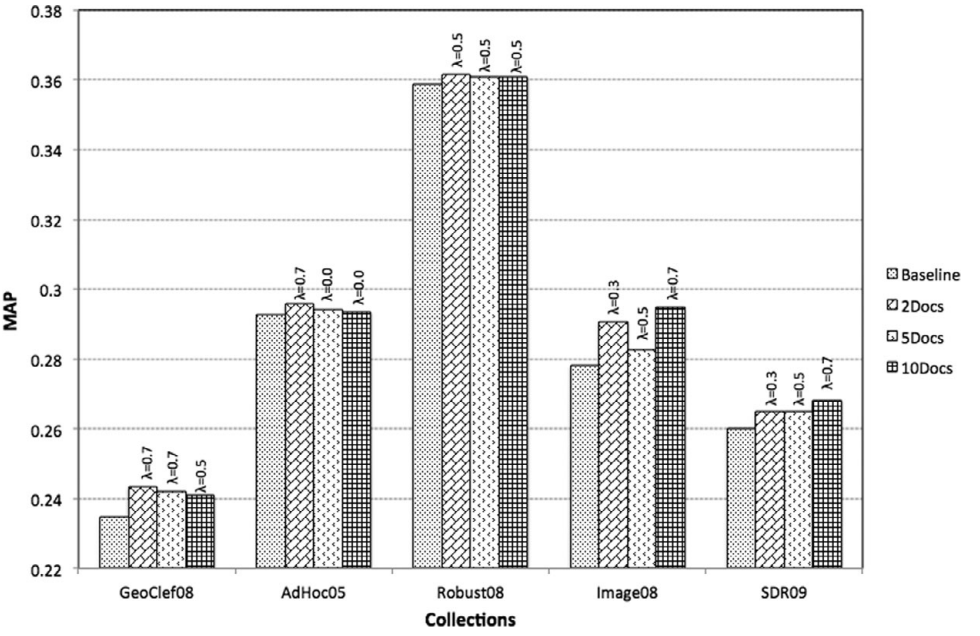


Fig. 4. Ranking Refinement results employing a blind feedback strategy. First, column represents the Base IR system performance; next, three columns show results obtained when 2, 5 and 10 *example documents* are automatically selected.

We considered two relevance feedback strategies for our experiments: (i) a blind scheme, where we took as *example texts* the first k retrieved documents by the base IR system; and (ii) a simulated scheme, where using ground truth[12] information, we simulate a manual relevance feedback session by identifying k *example texts* in the list. Section 6.1 describes the results obtained when a blind feedback scheme is adopted, and Section 6.2 the results when a simulated (*i.e.* manual) feedback is considered.

Experimental results are reported in terms of *MAP* values. Shown figures (Figures 4 and 5) summarise our results and expose only the best results obtained over the different configurations. In order to evaluate the statistical significance of the improvements that the MRF had over the baseline result, we performed the *paired Student's t-test*[13] with a confidence of 90 percent (i.e. $\alpha = 0.10$).

6.1 Experiments using blind relevance feedback

In this section, we show results obtained when a blind relevance feedback strategy is employed to select documents that will be used for defining the initial configuration of the proposed MRF, and for constructing the virtual document v. Figure 4 show

[12] Ground-truth information refers to the relevance judgements provided by the evaluation forum.

[13] This statistical test is traditionally employed in the field of IR since it is considered a more *appropriate* test for this task (Smucker, Allan and Carterette 2007).

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

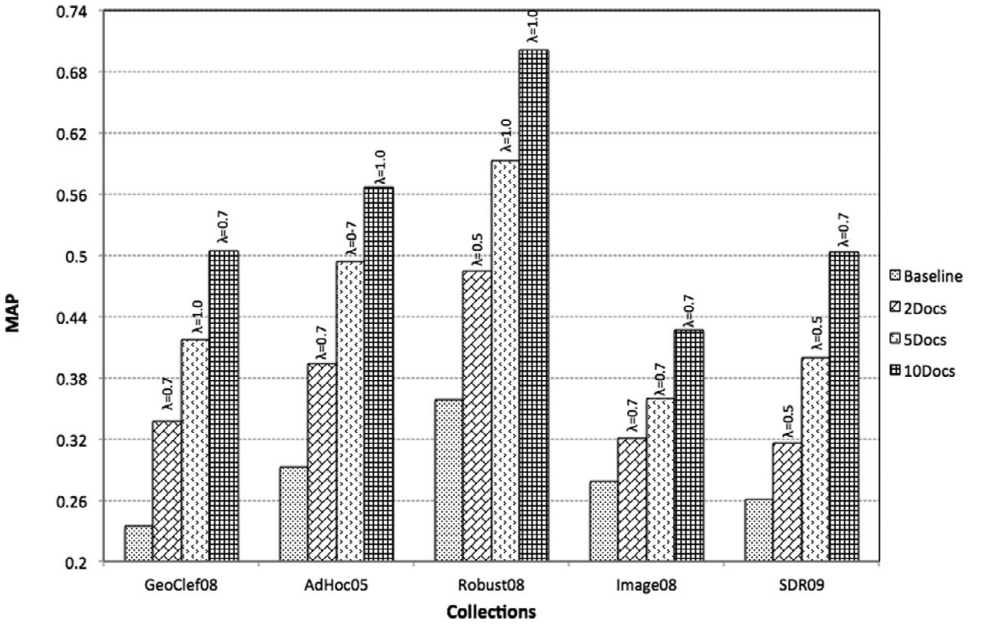


Fig. 5. Ranking Refinement results employing a simulated feedback strategy. First column represents the Base IR system performance; next three columns show results obtained when two, five and ten *example documents* are manually selected.

results obtained over the different considered tasks (See Section 5.3). The goal of this experiments was to quantify the improvement offered by the MRF for document re-ranking over an initial list of retrieved documents when there is no user intervention.

For the Geo08 task, it can be noticed that significant results are obtained when λ has values of 0.7 and 0.5. The best result occurs when two documents are selected as feedback and when λ has a value of 0.7. This means that for this collection the information provided by the neighbours has more benefit than the one provided by the observations. This configuration improves the base IR system up to a 3.71 per cent.

Similarly, to the Geo08 task, the best result for the AdHoc05 task is obtained when we employ two feedback elements and the scalar λ is equal to 0.7. However, this configuration does not represent a significant improvement; this means that even though it was possible to obtain the highest MAP value under this configuration, the improvement over all the queries was not consistent, which is also the case of λ equal to 0.0 and considering 10 documents as feedback. Nevertheless, when λ is equal to 0.0 considering five documents as feedback, our method is able to provide a ranking that is in fact a significant improvement.

It is worth mentioning that Robust08 task is considered a difficult task (See Table 2). However, better results are obtained when the scalar λ assigns equal importance to the *interaction* and *observation potentials*. Under this configuration and considering two *example documents* as feedback elements it is possible to improve the MAP in a 0.72 percent, showing all these cases significant improvements.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

In the case of the Image08 task, observe that one main difference of these results compared with previous experiments is that our MRF reach its best performance when 10 documents are considered as feedback, and similarly to previous experiments this happens when λ is 0.7, i.e. when more importance is given to the information provided by the neighbours. This configuration is able to significantly improve the base IR system by up to a 5.97 per cent in terms of the MAP measure.

Finally, for the SDR09 task, obtained results have a very similar behaviour than those obtained with the Image08 task. The best performance occurs when ten documents are considered as feedback and when the scalar $\lambda = 0.7$. In terms of the MAP measure, our MRF is able to significantly improve the base IR system by up to a 3.11 per cent.

In general, experiments showed that if a blind relevance feedback strategy is employed, it is recommended to use few feedback elements, since adding more feedback elements results in a degradation of the order assigned to the retrieved documents, which was the case for the "difficult" tasks (Geo08, AdHoc05 and Robust08). It is also important to mention that better results are obtained when the scalar λ get values between 0.3 and 0.7, indicating that the information provided by the neighbours and by the observation potential are both important to generate a better ordering of the documents.

If we know some characteristics of each of the evaluated tasks, it is also possible to assure that if our sets of queries are considered difficult (i.e. there are no many relevant elements within the document collection), it is recommended to use few feedback elements and to assign less importance to the *interaction potential*, and more relevance to the *observation potential*. On the other hand, if our set of queries are some how easier (i.e. are better supported by the collection), we can use more relevance feedback elements and trust more in the information provided by the neighbours (i.e. the interaction potential).

6.2 Experiments using simulated relevance feedback

In this section, we exhibit results obtained when a simulated relevance feedback is considered. As we previously mentioned, a simulated relevance feedback process consist of taking as feedback those documents that we know are true relevant elements (i.e. they are part of the ground truth). Similarly, to the experiments from previous section, we used these documents to initialise the MRF configuration, and to construct the virtual document v which will be part of the *observation potential*.

Figure 5 summarises obtained results for the tasks defined in Section 5.3. The main goal of this set of experiments is: (a) illustrate the possible reachable improvement of a base IR system if *minimal* user intervention is considered. It is worth mentioning that all results obtained under this feedback configuration were statistically significant with a confidence level of 99 per cent (i.e. $\alpha = 0.01$) in accordance with the *paired Student's t-test*.

Notice that for the Geo08 task, the best result considering either two or ten feedback elements is obtained when $\lambda = 0.7$. This is an indicator that both the

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

interaction potential and the *observation potential* are being complementary to each other. However, there still is a tendency to prefer higher values of λ (i.e. $\lambda \to 1$), which indicates that relevant neighbours have some homogeneity that allows the MRF to differentiate them from the irrelevant ones.

Similarly, in the case of the AdHoc05 task, better results with few feedback elements are reached when $\lambda = 0.7$. This configuration allows our MRF model to improve the MAP measure by up to a 34.59 per cent when only two documents are given as feedback.

Observe that for the Robust08 task, when only two documents are given as feedback elements, the best performance is obtained when $\lambda = 0.5$, which means that it is necessary for the MRF to complement information provided by the *interaction potential* (V_c) with the *observation potential* (V_a) information. However, if more elements are given as feedback it is possible to trust only in the information provided by the neighbours (i.e. the *interaction potential*).

Finally, for the case of the Image08 and SDR09 tasks, it is important to notice that both experiments reach their best performance when λ has values between 0.7 and 0.5. This indicates that for these particular collections, even if several documents are given as feedback elements, it is not possible to trust only in the information provided by the neighbours, which has been the case of the others tasks (Geo08, AdHoc05 and Robust08). In other words, the homogeneity among relevant documents within these collections (Image08 and SDR09) is not that evident, and as a consequence the MRF model needs of the information provided by the observations (i.e. the *observation potential*).

In general, performed experiments show that it is possible to outperform (consistently) results obtained by the base IR system. As expected, improvements are larger for higher values of k, nevertheless there are important improvements when $k = 2$, which is a very positive result since providing only two documents as examples (or feedback) is not an overwhelming task for the user.

Additionally, is was possible to observe that the proposed MRF behaves different depending on the characteristics of the document collection. For instance, Geo08, AdHoc05 and Robust08 use the same documents collection (See Table 1), which in general are news reports, and for this cases the MRF performs better when there is a tendency to prefer the information provided by the neighbours (i.e. $\lambda \to 1$). However, for Image08 and SDR09 which contain shorter documents, the MRF needs additional knowledge, i.e., the information provided by the observations in order to be able to provide a better ranking.

6.3 Performance analysis

6.3.1 On the selection of the parameters values

Previous experiments showed that the efficacy of the proposed MRF varies depending on the values for the parameter λ; which is a scalar that assigns more or less importance to our defined potentials (see expresion 5); and parameter k; which represents the number of *example documents*. In order to provide a deeper understanding of the impact of these parameters on the methods performance,

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

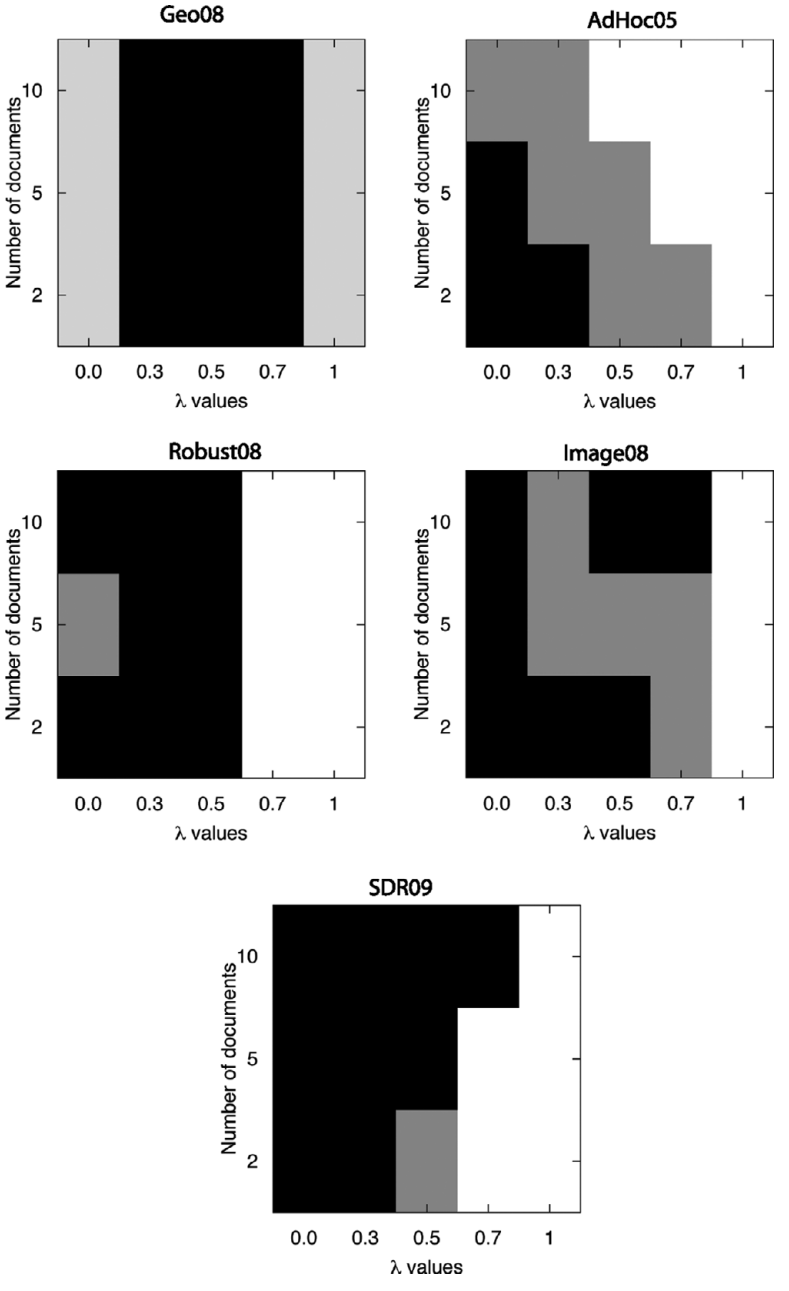


Fig. 6. Analysis of the statistical significance of results for several combinations of the configuration parameters of the proposed MRF employing a blind relevance feedback strategy.

Figure 6 plots the statistical significance of the improvements that the MRF – using a blind relevance feedback strategy – had over the baseline on the different collections, in accordance to the *paired Student's t-test* with a confidence of 90 per cent In these figures, a black dot indicates that the achieved improvement was

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

statistically significant, a gray dot indicates that the MRF was able to obtain a better performance in terms of the MAP measure but without being significant, whereas a white dot indicates that our method did not improve the results obtained by the base IR machine.

As can be observed, our method is able to generate a better ranking of the set of retrieved documents when both the interaction potential (V_c) and the observational potential (V_a) are considered in the energy function of the proposed MRF. Notice that in general, when the scalar λ is set to 1, it is not possible to obtain good results, indicating that the information provided by the neighbours is not enough to adequately find the best configuration of the MRF.

On the contrary, when λ goes from 0.0 to 0.7 it is possible to obtain results that represent a statistically significant improvement, particularly for λ set to 0.3. Also notice that by using few documents as *example texts* (two documents) is possible to obtain significant results. In general, we can conclude that the performance of our method was very robust across all evaluated tasks, suggesting that if a blind relevance feedback strategy is followed, it is recommendable to use few *example texts* and a scalar λ equal to 0.3.

For the experiments employing a simulated feedback process (Section 6.2), all configurations produce significant results, i.e. if we draw a similar graphs as the one shown in Figure 6 it will result in totally black graphs. Therefore, considering this fact and the results shown in Figure 5, we can conclude that by including only two *example texts* and considering an equilibrated weight between our potentials (V_c and V_a) it is possible to obtain significant improvements over the base system.

6.3.2 Running time efficiency

Although our goal was not to prove if the optimisation algorithm (ICM) produce better results in terms of solution quality and running time efficiency, we performed some experiments to measure the running time efficiency where the main variation was the number of documents that are taken into account for the construction of the fully connected MRF. For this experiments, we first consider a set of one hundred documents until we reach the 1,000 documents (considering increments of one hundred documents)[14].

Figure 7 presents the time efficiency of the proposed MRF when *example texts* are selected via a blind feedback strategy. As can be noticed, the greater the number of documents considered in the graph of the MRF, the greater the time required to converge to an optimal configuration.

Similarly, Figure 8 show the time efficiency of our method when the *example texts* are selected via simulated feedback. Notice that when *example texts* are provided by some user, the proposed MRF takes more time than its homologue experiment using blind feedback (Figure 7) to converge to the optimal configuration. The reason for this behaviour is because when a simulated feedback process is performed, we

[14] All of our experiments were run on Intel i7, 3.4GHz, 16GB RAM, and written in MATLAB R2008b version 7.7.0.471

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

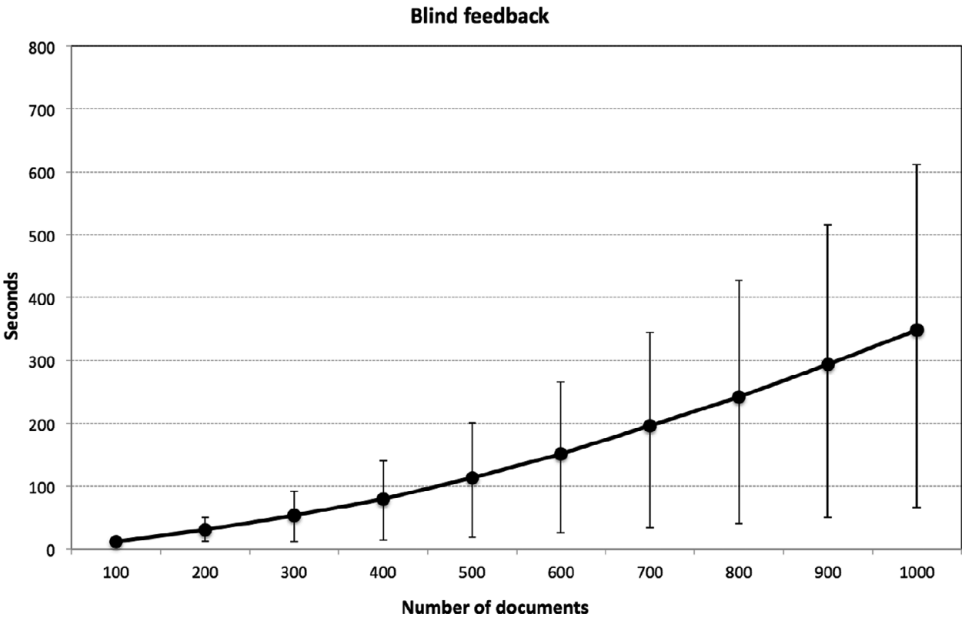


Fig. 7. Average time consumption for the blind feedback configurations with its respective standard deviations.

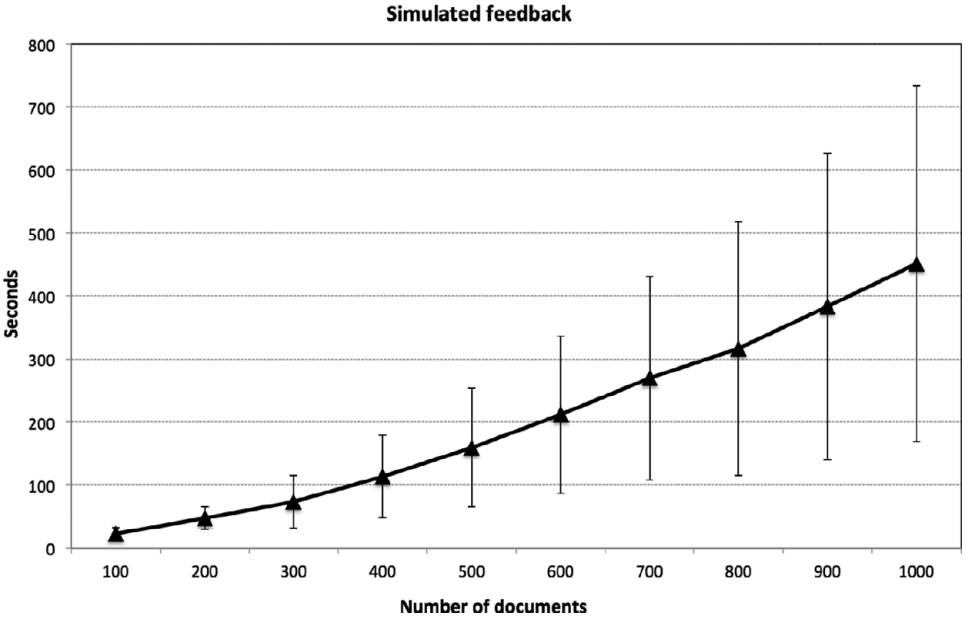


Fig. 8. Average time consumption for the simulated feedback configurations with its respective standard deviations.

are some how compelling the MRF to find documents with certain degree of homogeneity to the provided *example texts* that are in fact relevant documents, which is not always the case when a blind feedback process is performed.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

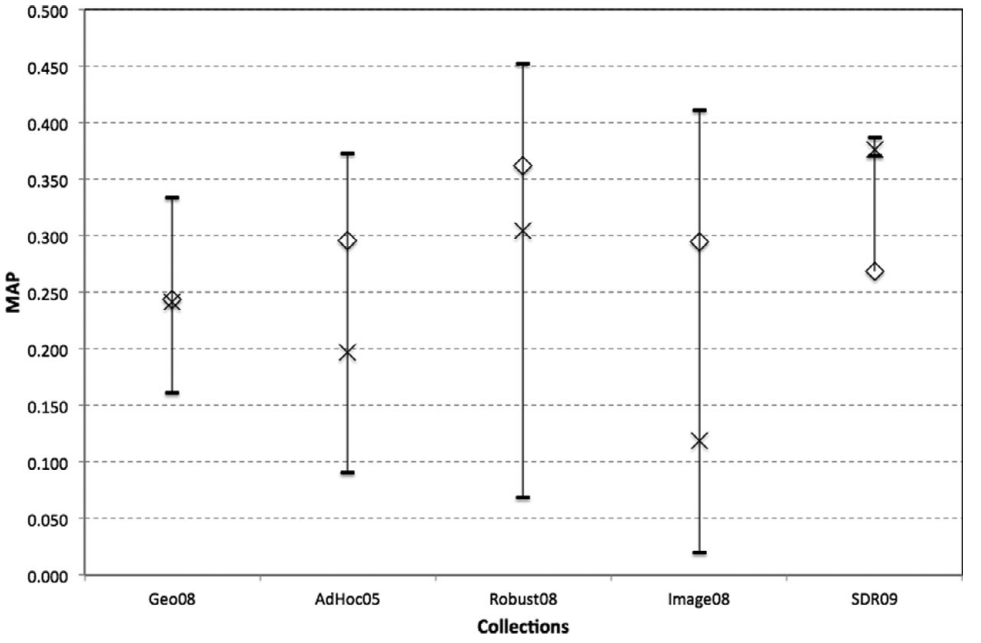


Fig. 9. Performance of the proposed method against the best obtained results in the bibliography. × represents the average performance of all participant teams. Upper line (-) indicates the maximum result obtained in each track, while the lower line (-) indicates the minimum performance. ◇ represents the performance obtained by our MRF using the blind feedback configuration (See Figure 4 for reference).

6.3.3 Comparison with other approaches

Finally, Figure 9 depicts a comparison of the performance obtained by the teams that participate in the most recent track for each task. Symbol × represents the average performance of all participant teams. Upper line (-) indicates the maximum performance obtained in each track, while the lower line (-) indicates the minimum performance[15]. Symbol ◇ represents the performance obtained by our proposed method using the blind feedback configuration (See Figure 4 for reference).

It is worth mentioning that the teams that participated in each track applied a great variety of methods and some times employed a great diversity of resources to solve the problem, which is the reason, we can not think in a direct comparison of our proposed method against them. However, Figure 9 demonstrates that our proposed method is able to reach a state-of-the-art performance except in the SDR09 task. It is also important to mention that when simulated feedback is used,

[15] These results for each task were taken from:
Geo08 (*http* : *//www.clef* − *campaign.org*/2008/*working_notes*/*AppendixE.pdf*)
AdHoc05 (*http* : *//www.clef* − *campaign.org*/2005/*working_notes*/*workingnotes*2005/*appendix_a.pdf*)
Robust08 (*http* : *//www.clef* − *campaign.org*/2008/*working_notes*/*AppendixC.pdf*)
Image08 (*http* : *//www.imageclef.org/system/files/ResultPublishing.xls.zip*)
SRD09 (*http* : *//trec.nist.gov/pubs/trec*9/*sdrt*9_*slides*/*index.htm*)

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

our method significantly outperformed the best results reported in the CLEF and TREC evaluation forums.

7 Additional experiments

In order to compare our findings against one of the most employed indirect ranking refinement strategies, i.e. QE via relevance feedback (QE), we conducted the following set of experiments, where the main goal was to: (i) compare our best results obtained with the proposed MRF model against QE results, (ii) to show the lack of sensitivity of our proposed MRF model to the quality of the information provided by the relevance feedback.

Each run of this set of experiments proceeds as follows. We took from the retrieved list by the base IR system *k example texts*, and all the information contained within these documents is employed to augment their respective query, and then a second retrieval process is performed using the expanded query. The new list of retrieved documents is then compared against the base IR system performance.

For the experiments presented in this section, we considered only a blind relevance feedback strategy since as we previously mention, we were interested in showing the robustness of our MRF method against the QE process when there is no user intervention.

7.1 QE via blind relevance feedback

In Figure 10, we present the MAP performance obtained for both the QE strategy and our ranking refinement (RR) method. For each evaluated task (i.e. Geo08, AdHoc05, Robust08, Image08 and SDR09), the first three bars represent results obtained with the QE process considering two, five and ten *example texts* respectively; and the following set of three bars represent the best results obtained employing our proposed ranking refinement strategy also considering two, five and ten *example texts*.

As it is possible to observe in Figure 10, QE technique is much more sensitive to the quality and to the quantity of added information to the original query, which is not the case of our proposed MRF method. Particularly, we can observe this behaviour for difficult tasks, which are Geo08 and Robust08. On the contrary, for better supported tasks (i.e. Image08 and SDR09), it can be sometimes preferable to perform a QE process.

In Figure 11, we exhibit the *recall* levels obtained by the QE process and the one of the base IR system, which is the same *recall* value maintained by our MRF method since as we have mentioned, we do not perform a second retrieval process. The first three columns correspond to the *recall* obtained using different configurations for the QE process, and the fourth column represents the *recall* value obtained by the base IR system.

Noticed that the QE process in general affects negatively the *recall* (see Figure 11). Hence, we can conclude that our MRF method for document ranking refinement is less sensible to the quality and to the quantity of added feedback elements, in other

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

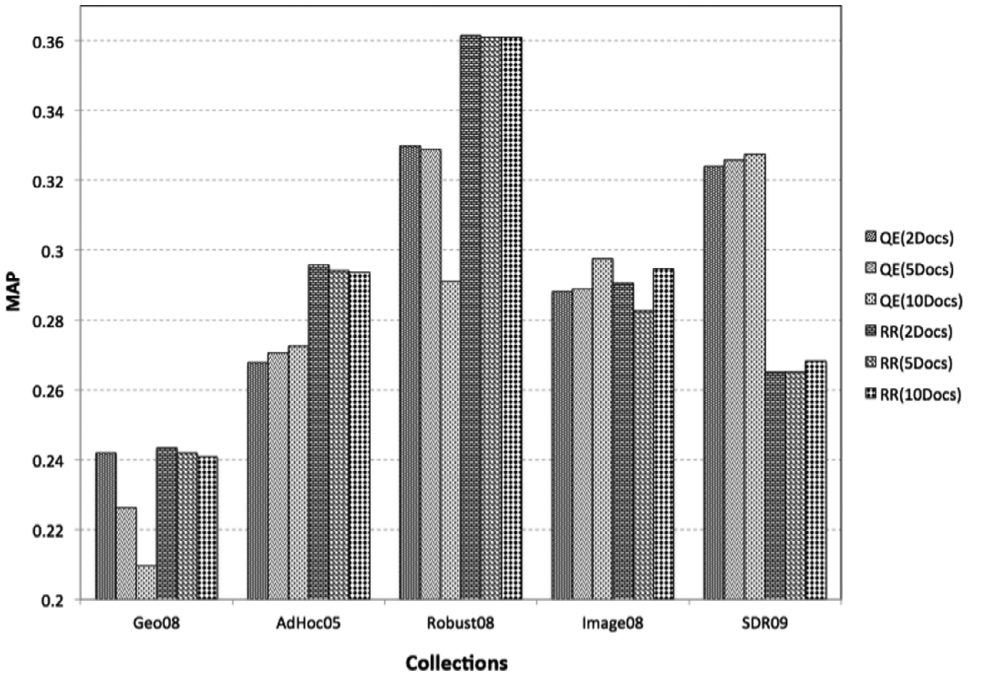


Fig. 10. Query Expansion *versus* Ranking Refinement using a MRF considering a blind feedback strategy. First three columns show the MAP performance of the QE technique using two, five and ten documents respectively. Last three columns indicate the MAP performance of the proposed MRF using two, five and ten *example documents* respectively.

words, the proposed MRF is a more robust method for this particular feedback situation where added information is partially correct.

8 Conclusions

We have introduced a MRF for improving the order of a list of initially retrieved documents. The proposed model incorporates similarity between documents in the list, similarity between documents and a set of *example texts*, and information obtained from the original order assigned to the documents. The ranking refinement problem is faced as one of separating relevant from irrelevant documents in the list. Our work included: (i) the inclusion of full texts (i.e. example texts) as feedback instead of a set of isolated terms, and (ii) the development of potentials and energy functions based on textual features that allow us to differentiate relevant from irrelevant documents.

Experimental results proved the effectiveness of our method. In particular, results showed that the proposed method is able to improve the base IR system. We studied the performance of the proposed model under different parameter settings and over several standard IR tasks from the CLEF and TREC forums. Experiments showed that under a blind relevance feedback scheme was possible to significantly improve results from a base IR system. These results also showed that few feedback elements are needed to reach a good performance. Additionally, we compare our method

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

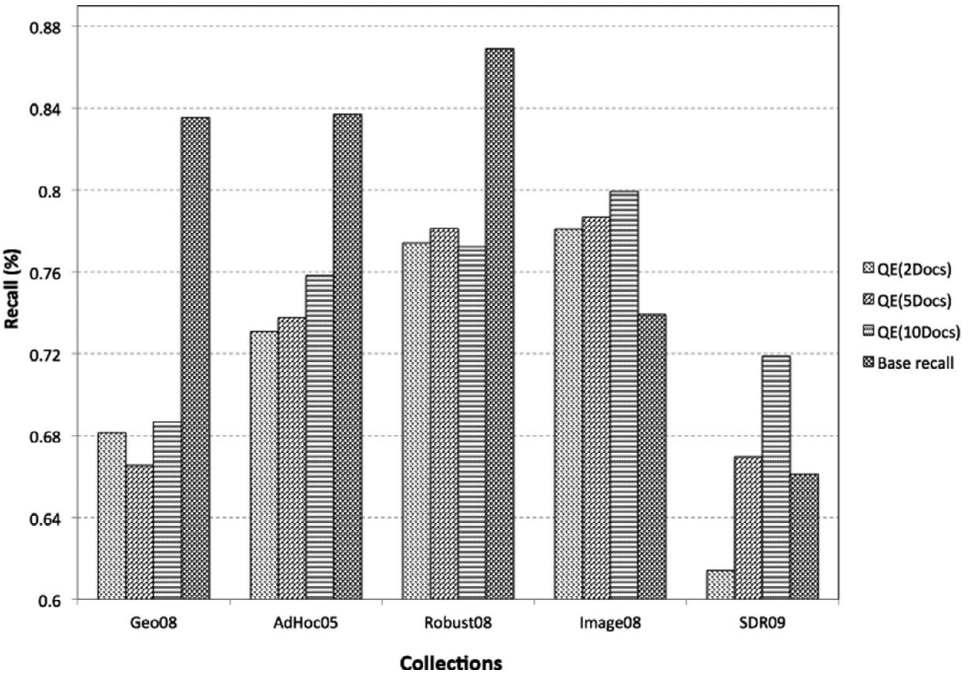


Fig. 11. Query Expansion *versus* Ranking Refinement using a MRF considering a blind feedback strategy. First three columns show *recall* levels obtained by the QE process considering two, five and ten documents respectively. Last column show the *recall* value obtained by the base IR system, which is the same *recall* value maintained by our MRF.

against one of the most employed ranking refinement strategies, i.e. QE. Obtained results showed that our MRF model is less sensitive to the quality and to the quantity of the feedback information than the QE technique.

The contributions of this work are as follows. We proposed a novel ranking refinement technique that faces the problem as one of combinatorial optimisation. The proposed model was able to improve the ranking of the base IR system as well as the performance of one of the most commonly used re-ranking strategies, i.e. a QE process. Our approach incorporates: (i) initial rank information provided by the base IR system; (ii) contextual information, which is often dismissed in usual post-retrieval techniques; and (iii) *example texts* information, which is obtained through a relevance feedback process. As observed in our experiments, our MRF model is capable of working with either automatic or manual relevance feedback strategies. Additionally, our MRF does not depend on any specific IR system, nor on a particular architecture or document collection.

As future work, we would like to include multimodal information to our MRF. For instance, in the Geo CLEF tasks besides considering textual features, we believe that it could be helpful to consider also geographical features. In the Image CLEF, this can be done by considering some visual features which can be obtained from the image provided as query, whereas for the SDR task this can be done by considering phonetic features. Additionally, we would like to consider in the energy function more

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

context information such as an user profile, or information of previous performed queries, which it will help to the MRF on determining potential relevant elements. Furthermore, we plan to implement and test our proposed method with a more recent and efficient energy minimisation algorithm, such as loopy belief propagation algorithm.

References

Baeza-Yates, R., and Ribeiro-Neto, B. 1999. *Modern Information Retrival*. Addison Wesley, Wokingham, UK.

Balinski, J., and Danilowicz, C. 2005. Re-ranking methos based on inter-document distance. *Information Processing and Management* **41**: 759–75.

Bear, J., Israel, D., Petit, J., and Martin, D. 1997. Using information extraction to improve document retrieval. In *Proceedings of the 6th Text Retrieval Conference*.

Bendersky, M., and Kurland, O. 2008. Re-ranking search results using document-passage graphs. In *Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR'08)*, pp. 853–4, ACM Press. Singapore, Singapore.

Besag, J. 1986. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B* **48**: 259–302.

Carbonetto, P., De Freitas, N., and Barnard, K. 2004. A statistical model for general context object recognition. In *Proceedings of the 8th European Conference on Computer Vision*, vol. 3021, pp. 350–62. Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Chávez, O., Sucar, L. E., and Montes, M. 2010. Image re-ranking based on relevance feedback combining internal and external similarities. In *Proceedings of The FLAIRS Conference*, Daytona Beach, Florida, USA.

Crouch, C., Crouch, D., Chen, Q., and Holtz, S. 2002. Improving the Retrieval Effectiveness of Very Short Queries. *Information Processing and Management* **38**.

Deng, H., Lyu, M. R., and King, I. 2009. Effective latent space graph-based re-ranking model with global consistency. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*, pp. 212–21, ACM Press. Barcelona, Spain.

Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. 2008. CLEF 2007: Ad Hoc Track overview. In *Post-proceedings of the 8th Workshop of the Cross Language Evaluation Forum CLEF 2007*, vol. 5152, pp. 13–32. Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Diaz, F. 2005. Regularising Ad Hoc retrieval scores. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*, pp. 672–79, ACM Press. Bremen, Germany.

Escalante, H. J., Montes, M., and Sucar, L. E. 2007. Word Co-occurrence and Markov random fields for improving automatic image annotation. In *Proceedings of the 18th British Machine Vision Conference*, vol. 2, pp. 600–9. Warwick, UK.

Garafolo, J. S., Auzanne, C. G. P., and Voorhees, E. M. 2000. The TREC spoken document retrieval track: a success story. In *Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access*, pp. 1–20, Paris.

Geman, S., and Geman, D. 1984. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. In *IEEE Transactions on: Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–41.

Grossman, D. A., and Frieder, O. 2004. *Information Retrieval, Algorithms and Heuristics*, 2nd ed. Springer. Dordrecht, The Netherlands.

Grubinger, M. 2007. *Analysis and Evaluation of Visual Information Systems Performance*. PhD thesis. School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University. Melbourne, Australia.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

Held, K., Kops, E., Krause, B., Wells, III, W., Kikinis, R., and Müeller, H. 1997. Markov random field segmentation of brain MR images. *IEEE Transactions on Medical Imaging* **16**: 878.

Hernández, C., and Sucar, L. E. 2007. Markov random fields and spatial information to improve automatic image annotation. In *Proceedings of the 2007 Pacific-Rim Symposium on Image and Video Technology*, vol. 4872, pp. 879–92. Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Kamps, J. 2004. Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In *Proceedings of the 21th European Conference on Information Retrieval*, vol. 2997, pp. 283–95. Lecture Notes in Computer Science. Springer.

Kemeny, J., Snell, J. L., and Kanpp, A. W. 1976. *Denumerable Markov Chains*. New York/Heidelberg/Berlin: Springer Verlag.

Kurland, O., and Lee, L. 2005. PageRank without hyper-links: structural re-ranking using links induced by language models. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*, pp. 306–13. ACM Press. Salvador, Brazil.

Lauritzen, S. L. 1996. *Graphical Models*. New York, NY: Oxford University Press.

Lease, M. 2009. An improved Markov random field model for supporting verbose queries. In *Proceedings of the 32th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, pp. 476–83. ACM Press. Boston, MA, USA.

Lee, K., Park, Y., and Choi, K. S. 2001. Document re-ranking model using clusters. *Information Processing and Management* **37**(1): 1–14.

Li, S. Z. 1994. Markov random field models in computer vision. In *Proceedings of the European Conference on Computer Vision*, vol. 18, pp. 361–70. Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Li, S. Z. 2001. *Markov Random Field Modeling in Image Analysis*, 2nd. ed. Springer.

Luk, R. W. P., and Wong, K. F. 2004. Pseudo-relevance feedback and title re-ranking for Chinese IR. In *Proceedings of the 4th NTCIR Workshop meeting*, Cross-lingual Information Retrierval Task. Tokyo, Japan.

Mallows, C. 1975. Non-null ranking models. *Biomedika* **44**: 114–30.

Metzler, D., and Croft, B. 2005. A Markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, pp. 472–9. ACM Press. Salvador, Brazil.

Metzler, D., and Croft, B. 2007. Latent concept expansion using Markov random fields. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, pp. 311–8. ACM Press. Amsterdam, The Netherlands.

Mitra, M., Singhal, A., and Buckley, C. 1998. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, pp. 206–14. ACM Press. Melbourne, Australia.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Mateo CA: Morgan Kaufman.

Porter, M. F. 1997. *An Algorithm for Suffix Stripping*, pp. 313–6. Morgan Kaufman Publishers Inc. San Francisco, CA, USA.

Qu, Y., Xu, G., and Wang, J. 2000. Rerank method based on individual thesaurus. In *Proceedings of the 2nd NTCIR Workshop on reserach in Chinese and Japanese Text Retrieval and Text Summarization*. Tokyo, Japan.

Salton, G., and Buckley, C. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science* **41**(4): 288–97.

Salton, G., Yang, C. S., and Wong, A. 1975. A vector space model for automatic indexing. *Communications of the ACM* **18**(11): 613–20.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press

Sarkar, P., and Moore, A. W. 2009. Fast dynamic reranking in large graphs. In *Proceedings of the 18th International conference on World Wide Web (WWW'09)*, pp. 31–40, ACM Press. Madrid, Spain.

Smucker, M. D., Allan, J., and Carterette, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*, pp. 623–32, ACM Press. Lisbon, Portugal.

Villatoro-Tello, E., Montes-y-Gómez, M., and Villaseñor-Pineda, L. 2009a. A ranking approach based on example texts for geographic information retrieval. In *Post-Proceedings of the 9th Workshop of the Cross Language Evaluation Forum CLEF 2008,* vol. 5822, pp. 239–50. Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Villatoro-Tello, E., Villaseñor-Pineda, L., and Montes-y-Gómez, M. 2009b. Ranking refinement via relevance feedback in geographic information retrieval. In *Proceeding of the Mexican International Conference on Artificial Intelligence MICAI 2009*, vol. 5845, pp. 165–76. Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Winkler, G. 2006. Image analysis, random fields and Markov chain monte carlo methods. *Springer Series on Applications of Mathematics*, Rozovskii B. and Yor M. eds. Vol. 27, pp. 179–96, 2nd ed. Springer, Germany.

Xu, J., and Croft, W. B. 1996. Query expansion using local and global document analysis. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4–11. ACM Press. Zurich, Switzerland.

Xu, J. and Croft, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems* **18**(1): 79–112.

Yang, L. P., and Ji, D. H. 2005a. Chinese information retrieval based on terms and relevant terms. *ACM Transactions on Asian Language Information Processing* **4**(3): 357–74.

Yang, L. P., and Ji, D. H. 2005b. Chinese document re-ranking based on term distribution and maximal marginal relevance. In *Proceedings of the 2nd Asian Information Retrieval Symposium AIRS*, vol. 3689, pp. 299–311. Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Yang, L., Ji, D., and Tang, L. 2004. Document re-ranking based on automatically acquired key terms in Chinese information retrieval. In *COLING '04 Proceedings of the 20th International Conference on Computational Linguistics*. pp. 480–6. Association for Computational Linguistics. Geneva, Switzerland.

Yang, L., Ji, D., Zhou, G., Nie, Y., and Xiao, G. 2006. Document re-ranking using cluster validation and label propagation. In *Proceedings of the ACM CIKM 2006 International Conference on Information and Knowledge Management*, pp. 690–7. ACM Press. Arlington, Virginia, USA.

Zhang, B., Hua, L., Yi, L., Lei, J., Wensi, X., Weiguo, F., Zheng, C., and Wei-Ying, M. 2005. Improving web search results using affinity graph. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 504–11. ACM Press. Salvador, Brazil.

Zhou, D., Lawless, S., Min, J., and Wade, V. 2010a. A late fusion approach to cross-lingual document re-ranking. In *Proceedings of the ACM CIKM 2010 International Conference on Information and Knowledge Management*, pp. 1433–6. ACM Press. Toronto, ON, Canada.

Zhou, D., Lawless, S., Min, J., and Wade, V. 2010b. Dual-space re-ranking model for document retrieval. In *COLING '10 Proceedings of the 23rd international conference on Computational Linguistics*, pp. 1524–32. Association for Computational Linguistics. Beijing, China.

https://doi.org/10.1017/S1351324912000010 Published online by Cambridge University Press