Analogy-based Matching Model for Domain-specific Information Retrieval

Myriam Bounhas[1,2] and Bilel Elayeb[1,3]
[1]*Emirates College of Technology, Abu Dhabi, United Arab Emirates*
[2]*LARODEC Research Laboratory, ISG of Tunis, Tunis University, Tunisia*
[3]*RIADI Research Laboratory, ENSI, Manouba University, Tunisia*

Keywords: Information Retrieval, Analogical Proportions, Similarity, Agreement, Disagreement, Analogical Relevance.

Abstract: This paper describes a new matching model based on analogical proportions useful for domain-specific Information Retrieval (IR). We first formalize the relationship between documents terms and query terms through analogical proportions and we propose a new analogical inference to evaluate document relevance for a given query. Then we define the *analogical relevance* of a document in the collection by aggregating two scores: the Agreement, measured by the number of common terms, and the Disagreement, measured by the number of different terms. The disagreement degree is useful to filter documents out from the response (retrieved documents), while the agreement score is convenient for document relevance confirmation. Experiments carried out on three IR Glasgow test collections highlight the effectiveness of the model if compared to the known efficient Okapi IR model.

1 INTRODUCTION

Reasoning by analogy (Prade and Richard, 2010), which straddles the fields of artificial intelligence and linguistics, is the basis of psychological foundations of human behaviour, in which an inference is applied to analyze and categorize new problems by highlighting their resemblance to problems already solved. Analogical proportions are recognized as useful tools to perform comparisons between situations expressed by differences that are equated to other differences. More precisely, they are statements of the form: *x is to y as z is to t*, often denoted *x : y :: z : t* that express that "x differs from y as z differs from t", as well as "y differs from x as t differs from z" (Miclet and Prade, 2009). In terms of pairs, we can consider that the pair *(x,y)* is analogous to the pair *(z,t)* (Hesse, 1959). Analogical proportions are based on the assumption that if four objects *x*, *y*, *z*, *t* are making an analogical proportion on a set of given features, it may also continue holding on another sign related to them. The problem is then to predict this sign for *t* based on the known signs for *x*, *y* and *z* in case the signs make an analogical proportion.

Analogical proportions have been recognized as an interesting direction in the last two decades (Lepage, 2001; Yvon et al., 2004; Stroppa and Yvon, 2005b; Miclet and Prade, 2009; Prade and Richard, 2013). They have demonstrated their ability to provide operational and effective models for morphological linguistic analysis (Stroppa and Yvon, 2005a) and classification tasks developed first by (Bayoudh et al., 2007; Miclet et al., 2008) and extended by (Prade et al., 2012) and (Bounhas et al., 2017a) and has led to encouraging results in terms of accuracy and complexity. In classification case, the predicted sign is the label of the class.

IR problems are founded on the idea of assigning relevant documents for a given query. It is natural to think that dissimilar queries should lead to a very distinct set of relevant documents. In contrary, the set of relevant documents for too similar queries should not be distinguishable. From an analogical point of view, this is can be expressed as a matter of comparisons between queries and the corresponding set of relevant documents. Starting from this assumption, this leads us to wonder if what is successfully working in classification may be applied to information retrieval. The problem in this case is no longer to predict the class for a new example but rather the relevance of a document to a new query.

In this paper, we mainly focus on information retrieval based on the idea of analogy between queries and documents and we propose a new Analogical

Bounhas, M. and Elayeb, B.
Analogy-based Matching Model for Domain-specific Information Retrieval.
DOI: 10.5220/0007342104960505
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 496-505
ISBN: 978-989-758-350-6

Proportion-based Matching Model (APMM). Given two queries q_1 and q_2 with their corresponding set of relevant documents $rel(q_1)$ and $rel(q_2)$, this model assumes that q_1 differs from q_2 as $rel(q_1)$ differs from $rel(q_2)$. This means that the difference between the two queries may considerably affect the difference between their corresponding relevant documents. Based on this logic and given a new query, the idea of *APMM* is to *guess* the relevance/irrelevance of any document in the collection to this new query, based on other existing queries. The matching model that we propose is dedicated for domain-specific information retrieval as a sub-domain of IR in which queries and document collection are issued from a specific domain, such as social science, medical information, aeronautic, etc. Given a domain-specific IR test collection, we exploit the idea of analogy between a "test" query (new query), "training queries" (past queries) along with their corresponding documents to search for relevant documents to this new test query.

This paper is organized as follows. Section 2 recalls the basic definitions and properties of analogical proportions. In Section 3, we describe how to exploit analogical proportions to evaluate document relevance. For this purpose, we propose new Agreement and Disagreement scores. This forms the basis for defining the final Analogical Relevance measure for documents. In Section 4, we present the analogy-based IR matching model and we propose a new algorithm for this purpose. Section 5 details the experiments carried out on three IR Glasgow test collections and compares the effectiveness of the model to the known efficient Okapi IR model.

2 BACKGROUND ON ANALOGICAL PROPORTIONS

As already said, *an Analogical proportion* denoted as $a : b :: c : d$, can be read: a **is to** b **as** c **is to** d or more informally as "a differs from b as c differs from d and vice versa". It is considered as a special case of logical proportion and defined as (Prade and Richard, 2012):

$$a : b :: c : d = (a \wedge \neg b \equiv c \wedge \neg d) \wedge (\neg a \wedge b \equiv \neg c \wedge d) \tag{1}$$

Analogical proportion satisfies diverse properties, usually expected from numerical proportion such as:

- *Symmetry*: $a : b :: c : d \Rightarrow c : d :: a : b$ and
- *Central permutation*: $a : b :: c : d \Rightarrow a : c :: b : d$.
- Thanks to the central permutation property, a third property requires that the two following implications also hold: $a : b :: a : x \Rightarrow x = b$ and $a : a :: b : x \Rightarrow x = b$
- *Transitivity*: $a : b :: c : d \wedge c : d :: e : f \Rightarrow a : b :: e : f$.

Let u, v be two distinct values in a finite set U, an analogical proportion always holds for the three following patterns: (u,u,u,u), (u,u,v,v) and (u,v,u,v). All other possible patterns with two distinct values disagree with the idea of analogical proportions. More precisely, (u,u,u,v), (u,u,v,u), (u,v,u,u), (v,u,u,u), and (u,v,v,u) are invalid patterns. In fact, assuming that "u is to u as u is to v" for $u \neq v$ seems strange.

The above definition of analogical proportions can easily be extended to items represented as *vectors* of values. In the Boolean setting, let S be a set of vectors $\in \{0,1\}^n$, each vector $\overrightarrow{x} \in S$ is represented by n features as $\overrightarrow{x} = (x_1, \cdots, x_n)$. Given four vectors $\overrightarrow{a}$, $\overrightarrow{b}$, $\overrightarrow{c}$ and $\overrightarrow{d} \in S$. For each feature $i \in [1,n]$, there are only eight possible combinations of values of $\overrightarrow{a}$, $\overrightarrow{b}$ and $\overrightarrow{c}$ (see Table 1). We can see that in two situations among eight, the equation can not be solved.

Table 1: Solving analogical proportion for Boolean vectors.

$\overrightarrow{a}$	0	0	0	0	1	1	1	1
$\overrightarrow{b}$	0	0	1	1	0	0	1	1
$\overrightarrow{c}$	0	1	0	1	0	1	0	1
$\overrightarrow{d}$	0	1	1	?	?	0	0	1

Let us consider the analogical equation $\overrightarrow{a} : \overrightarrow{b} :: \overrightarrow{c} : \overrightarrow{d}$ between four Boolean vectors. To solve this equation, it is common to apply the following extension of the previous definitions to Boolean vectors in $\{0,1\}^n$:

$$\overrightarrow{a} : \overrightarrow{b} :: \overrightarrow{c} : \overrightarrow{d} \text{ iff } \forall i \in [1,n],\ a_i : b_i :: c_i : d_i$$

which supposes that analogical proportion between four vectors holds true iff the analogical proportion holds componentwise between *all* their features.

The above equation solving property forms the basic to define an inference principle for binary classification problems applied to Boolean datasets in (Bayoudh et al., 2007; Bounhas et al., 2017a). Based on the continuity principle, the authors assumed that if the analogical equation holds componentwise for *all* features of four Boolean instances, this analogical equation should still holds for their classes. Having four Boolean instances $\overrightarrow{a}$, $\overrightarrow{b}$, $\overrightarrow{c}$ and $\overrightarrow{d}$, the first three instances are in the training set with known classes $cl(\overrightarrow{a})$, $cl(\overrightarrow{b})$, $cl(\overrightarrow{c})$ and the last one whose class is unknown (to be classified). The inference principle is defined as:

$$\frac{\forall i \in [1,n], a_i : b_i :: c_i : d_i}{cl(\overrightarrow{a}) : cl(\overrightarrow{b}) :: cl(\overrightarrow{c}) : cl(\overrightarrow{d})}$$

To classify the new instance $\overrightarrow{d}$, the equation $cl(\overrightarrow{a}) : cl(\overrightarrow{b}) :: cl(\overrightarrow{c}) : x$ should be solvable and then assign its solution to $cl(\overrightarrow{d})$.

3 EVALUATING DOCUMENT RELEVANCE USING ANALOGICAL INFERENCE

In this section, we investigate the ability of analogical proportions to model the relationship between queries and documents. Let us consider a set of n queries $Q = (q_1, q_2, ..., q_n)$ and their corresponding sets of relevant documents $D = (rel(q_1), rel(q_2), ..., rel(q_n))$, s.t: $rel(q_i) = \{d_1, d_2, ..., d_w\}$ is the set of relevant documents for a query q_i.

Based on analogical proportions, the "query-document" relationship may have a new meaning based on linking pairs of queries. Namely, q_i differs from q_j as $rel(q_i)$ differs from $rel(q_j)$. More precisely, the extent to which queries q_i and q_j are similar/dissimilar should *strongly affect* the identity/difference of sets $rel(q_i)$ and $rel(q_j)$ in terms of the relevance or irrelevance of each document from the collection D. Given a new query (unseen before), the basic idea is to guess the relevance/irrelevance of any document $d \in D$ to this new query based on existing queries.

In order to understand better the above idea, we need to represent in a more precise way the link between queries and documents. We assume here that both queries and documents are indexed in the same way and can be represented through a set of indexing terms. Let us consider two queries q_i, q_j and a document d_k denoted as:

$$q_i = (t_1^i, t_2^i, ..., t_p^i), q_j = (t_1^j, t_2^j, ..., t_{p'}^j) \text{ and}$$

$$d_k = (t_1^k, t_2^k, ..., t_{p''}^k)$$

If we consider a particular term t, we define a predicate $q(t)$(resp. $d(t)$) as a boolean value which is equal to 1 if the term t exists in the query q (resp. document d) and 0 otherwise. We also define the predicate $rel_{ik}(t)$ which is evaluated to 1 if, according to the term t, the document d_k is relevant to the query q_i and 0 otherwise (we assume here that each term t provides us a bit of knowledge about the relevance or irrelevance of this document). Following this logic, one may represent query-document relationship using analogical proportion in two different ways:

- (i) *Term existence*: In a first level, we aim to represent the existence or non existence of a particular term in a query or document. For this purpose, we propose to consider the following analogical proportion:

$$q_i(t) : q_j(t) :: d_k(t) : q_j(t) \qquad (2)$$

This proportion states that the term t exists/not exists in queries q_i and q_j in the same way as it exists /not exists in document d_k and query q_j. This means that, for a particular term t, the difference between two queries is the same as the difference between a document and one of the two queries.

- (ii) *Document relevance*: In a second level, we wonder about the relevance/irrelevance of a document for a given query. The analogical proportion:

$$q_i(t) : q_j(t) :: rel_{ik}(t) : rel_{jk}(t) \qquad (3)$$

states that the difference in terms of existence/non-existence of a term between two queries and a document implies the difference in terms of relevance/irrrelevance of the given document, satisfying term existence proportion (eq. 2), for these queries.

It is clear that, the second analogical proportion (eq. 3) can only be applied if the first one holds. In fact, we assume here that the existence/non existence of a term in a query and a document may be a good indicator of the relevance/irrelevance of this document to this query, which means that we may induce document relevance from the existence of the term in both query and document. This leads to the following analogical inference:

$$\frac{q_i(t) : q_j(t) :: d_k(t) : q_j(t)}{q_i(t) : q_j(t) :: rel_{ik}(t) : rel_{jk}(t)}$$

Since in IR, queries and documents are previously indexed with terms in a preliminary step, the first type of analogical proportion (eq. (2)) is known from the training set. If this first equation holds, it helps to infer the second type of analogical proportion (eq.(3)) which helps to induce document relevance.

The analogical inference that we propose seems close to that stated first in classification problems (Bayoudh et al., 2007). However, the logic is different: in classification, if the analogical equation between the four components holds for *all* features, the classification inference is applied to guess the final class of the instance to be classified. In our model, since we treat indexing terms independently, the above analogical inference is applied for *each* term in the query to induce document relevance *according* to this particular term. Then, the final document relevance is guessed by aggregating *individual* document relevance induced from each term in the query. This will be explained in the next subsections.

Table 2: Relevance truth values.

	$q_r(t)$	$q_x(t)$	$d_k(t)$	rel_{rk}	rel_{xk}	
(i)	**1**	**1**	**1**	$-$		
(ii)	**1**	**1**	1	**1**	**y= 1**	$\|A_t$
(i)	**1**	**0**	**1**	$-$		
(ii)	**1**	**0**	1	**1**	**y=0**	$\|D_t$

The search process for relevant documents is based on the resolution of analogical equations described above. This process assumes that: if the two queries q_i and q_j are in analogical proportion (with some documents d), it should be the case for their corresponding relevance/irrelevance predicates with regard to the same document. In next subsection, we describe how this inference process can be applied to evaluate the relevance /irrelevance of each document in the collection.

3.1 Agreement and Disagreement Scores

Let us consider the query-document matching problem in IR where a query q_r is in the training set having its known corresponding relevant document set $rel(q_r) = \{d_1^r, d_2^r, ..., d_w^r\}$. A second query q_x is extracted from the test set and whose $rel(q_x)$ is unknown. Starting from the training set, the equation $q_r(t) : q_x(t) :: d_k(t) : q_x(t)$ is solvable. To build the set of relevant documents $rel(q_x)$ for the new query q_x, one may start by looking at each document $d_k \in rel(q_r)$ for each known query q_r. If the analogical proportion given in equation (3): $q_r(t) : q_x(t) :: rel_{rk}(t) : y$ has a solution, we assign to $rel_{xk}(t)$ its solution for each document $d_k \in rel(q_r)$. In the following, we first analyze the truth table for the two proposed equations (2) and (3) introduced before. Then, we propose an agreement/disagreement scores that help to evaluate document relevance. Table 2 provides the truth values of the predicate rel_{xk} as a solution of equation (3). In each line of this Table, only *bold* truth values are to be considered appropriately to solve each analogical eq.(2) and (3).

To solve the above analogical equation, it is clear that *only* two different situations (see Table 2) are important to consider. In terms of generic patterns, we can see that the analogical proportion (2), for example, always holds for the two following patterns: *u*:*u*::*u*:*u* (two first rows in Table 2) and *u*:*v*::*u*:*v* (two last rows) where *u* and *v* are distinct values. However and as introduced in Section 2, analogy should also hold for the pattern *u*:*u*::*v*:*v*. Nevertheless, this situation is meaningless in our inference process since in equation (2), the second and last predicates are assumed to be the same. Moreover, for the two used patterns we only consider the case where the predicate $q_r(t)$ (or *u*) is true since when it is false it does not help for predicting the relevance of a document: we focus on terms existing in the query q_r not on those that are not existing.

Based on the patterns in Section 2, there are two indicators of document relevance: The pattern *u*:*u*::*u*:*u* states for a total agreement between the two queries and the document according to the existence of term t. This agreement should also be applied to guess document relevance in equation (3) and thus may be considered as a good indicator that the studied document is also *relevant* for q_x (i.e: $y = 1$). In contrary, the pattern *u*:*v*::*u*:*v* states for a disagreement between the two queries according to term existence (eq. (2)). Applying this disagreement in the same way to equation (3) reinforce the idea that this document is rather *irrelevant* for q_x (i.e: $y = 0$). We define the agreement and disagreement scores by:

- *Term Agreement*: According to term t, a document d_k is *relevant* for a test query q_x if this query *agree* with both query q_r and its relevant document d_k on the existence of this term.
- *Term Disagreement*: According to term t, a document d_k is *irrelevant* for a test query q_x if this query do *not agree* with query q_r and with its relevant document d_k on the existence of this term.

Let us now state the previous ideas with formal notations. Given a particular term t, in the following we define two scores A_t and D_t to evaluate the extent to which a query q_x is in $agreement/diagreement$ with another query q_r, and a document d_k as:

$$A_t(q_r, q_x, d_k) = q_r(t) \wedge q_x(t) \wedge d_k(t) \wedge rel_{rk}(t) \quad (4)$$

$$D_t(q_r, q_x, d_k) = q_r(t) \wedge \neg q_x(t) \wedge d_k(t) \wedge rel_{rk}(t) \quad (5)$$

We can easily check that A_t and D_t appropriately rewrite the analogical proportions (2) and (3). In fact, we assign to $rel_{xk} = A_t \wedge \neg D_t$ the solution of $(q_r(t) : q_x(t) :: d_k(t) : q_x(t)) \wedge (q_r(t) : q_x(t) :: rel_{rk}(t) : y)$. The two proposed scores will be used to define a global agreement/disagreement scores between queries and documents with regard to *all* terms.

3.2 Global Agreement and Disagreement Scores

It is known in IR that queries and documents can simply be represented by a set of indexing terms. The indexing process of queries and documents can be done in a preliminary step before applying the matching model. Starting from the previously defined agreement/disagreement scores related to one term t, we have to define a global score suitable for a set of terms. The agreement and disagreement scores defined by equations (4) and (5) are used to evaluate the extent to which a test query agrees or disagrees with other seen queries and their corresponding relevant documents if we only look to one particular term t. Given a set of terms for each query and assuming independence of indexing terms, one may estimate the global agreement/disagreement of a query q_x with a query $q_r = (t_1, t_2, ..., t_m)$ as the *sum* of agreement/disagreement of each term $t_i, i \in [1, m]$ as follows:

$$Ag(q_r, q_x, d_k) = \frac{1}{m} \sum_{i=1}^{m} A_{t_i}(q_r, q_x, d_k)$$

$$Dis(q_r, q_x, d_k) = \frac{1}{m} \sum_{i=1}^{m} D_{t_i}(q_r, q_x, d_k)$$

where m is the number of terms in q_r. For a given test query q_x, the matching model aims to evaluate each distinct pair (query, document) i.e: (q_r, d_k), where q_r is in the training set and d_k is among its relevant documents. This enables to select relevant documents for q_x among those relevant for other seen queries. Combining each pair (q_r, d_k) with q_x forms a set of triples (q_r, q_x, d_k). To compute the agreement/disagreement for each triple in this set, we can estimate $Ag(q_r, q_x, d_k)$ for the set of terms where they agree (i.e. the term t exists) and $Dis(q_r, q_x, d_k)$ for the set of terms where they disagree (i.e. the term t exists in q_r and d_k and does not exist in q_x).

To understand better this idea, in Table 3 we represent term predicates for each element of the triple to show the Ag/Dis situations.

Table 3: Ag/Dis scores with regard to *all* indexing terms.

	Ag			**Dis**			
	t_1	...	t_{k-1}	t_k	...	t_m	*rel*
q_r	1	...	1	1	...	1	
q_x	1	...	1	0	...	0	
d_k	1	...	1	1	...	1	$rel_{xk} =?$

After reordering q_r terms in Table 3, we can see that q_r and q_x agree on terms t_1 to t_{k-1} and disagree on terms t_k to t_m. Consider now the document d_k known to be relevant for q_r and agree with q_x in the same way as q_r. It is clear that the equations (2) and (3) hold componentwise between the three elements of the triple. The aim is thus to guess the relevance of document d_k for q_x based on the amount of agreement/ disagreement terms with both q_r and q_x.

It is natural to consider a document d_k (known to be relevant for q_r) as likely to be also relevant for a query q_x if it contains as much as agreement terms and no disagreement terms. To select the best documents with high relevance, it is recommended to choose those having high value of Ag and small value of Dis. High value of Ag means that for large number of terms, the document agrees with the test query. On the opposite, small value of Dis guarantee a reduced number of terms for which the document disagrees with the test query. In the very optimistic case, one wants a document d_k to agree with the query q_x with respect to *all* terms (Ag close to 1) and to disagree on no term (Dis close to 0). We define the *Analogical Relevance* of a document d_k for a query q_x as:

$$AR(q_r, q_x, d_k) = min(Ag(q_r, q_x, d_k), 1 - Dis(q_r, q_x, d_k)) \tag{6}$$

It is natural to assume that any document d_k in the collection D may be relevant for different training queries at the same time ($d_k \in rel(q_r)$ and $d_k \in rel(q'_r)$ with $q_r \neq q'_r$). This means that, for each candidate document d_k, we have to evaluate its analogical relevance AR with regard to each training query q_r where $d_k \in rel(q_r)$. Then, the final *Analogical Relevance* of document d_k is obtained by aggregating all these AR's evaluations on all training queries using the *max*:

$$AnalogicalRelevance(q_x, d_k) = \max_{r=1}^{n}(AR(q_r, q_x, d_k)) \tag{7}$$

The analogy-based matching model that we propose *combines* the two scores of agrement and disagreement to guess document relevance. In fact, the disagreement indicator, measured by the number of different terms, removes from the list of returned documents, for a given query, those that are not relevant, while the agreement indicator, measured by the number of common terms, reinforces the relevance of the remaining documents which are not eliminated by the disagreement indicator. The proposed matching model has its counter part in classification problems. In fact, the analogy-based classification approach of Bounhas et al. (Bounhas et al., 2014) treats the similarity and dissimilarity *sequently* in separate levels. In their approach, they first look for the most similar pairs (a, b) having the maximum number of similar attributes. This enables to filter the candidate voters for classes. Then check the analogical equation on the remaining dissimilar attributes between pairs (a, b) and (c, d) to solve the analogical equation for classes. The authors have proven the efficiency of this approach to reduce the average number of used triples for classification.

4 ANALOGICAL PROPORTION-BASED MATCHING MODEL

In this section, we detail the proposed matching model, denoted here *APMM*, based on the previously defined *Analogical Relevance* measure.

4.1 Methodology

Let $Q = (q_r, rel(q_r))$ be a training set of queries with their known corresponding relevant documents. Given a new query $q_x \notin Q$, a first option to retrieve its relevant documents is to start by a force brute method in which all queries q_r in the training set Q along with their corresponding relevant documents are considered in the search process. Each document in each pair (q_r, d_k) (where $d_k \in rel(q_r)$) is assumed to be a *candidate relevant document* for q_x. Retrieved documents are simply the solution of the equation (3). Based on the *analogical relevance* function defined by equation (7), we first compute $AnalogicalRelevance(q_x, d_k)$ for each document, then choose the best ones as a final set of relevant documents for q_x. As can be seen, the complexity of this force brute approach is quadratic due to the search space for pairs of (query, document). This process may become time consuming for large number of queries and/or relevant documents for each query. To optimize this first approach, we have chosen to use only the *nearest neighbors* queries to the new query q_x in the search process of relevant documents. This will considerably reduce the search space to be linear. In fact, we wonder if the study of the set of relevant documents of the nearest neighbors queries q_r to q_x are enough to retrieve the most relevant documents to this new query. Similar approaches, based on the idea of nearest neighbors applied to classification, has been developed and achieved successful results for Boolean or numerical data (Bounhas et al., 2017b; Bounhas et al., 2018).

In practice, our implementation can be summarized by the following steps:

- Given a new query q_x in the test set, find its k nearest neighbors queries in the training set $q_r \in Q$.
- Build the set of candidate relevant documents. This set is simply the union of *all* relevant documents corresponding to the k-nearest neighbors queries.
- In a filtering step, compute the $AnalogicalRelevance$ for each candidate document then remove those whose $AnalogicalRelevance(q_x, d_k)$ is less than a fixed threshold.

The proposed matching model can be described by the following algorithm.

4.2 Algorithm

Let $CosineSim(q_x, q_r)$ be a function that returns the *cosine similarity* between the two queries q_x and q_r (Amit, 2001). This function enables to order training queries q_r according to their similarity to the test query and then select the k-nearest neighbors (k is a given value). The computation of the $NN_k(q_x)$'s can be done offline in a pre-processing step, which helps to speed up the matching process. The previous explanation can now be described with Algorithm 1. It is important to know that, contrary to other matching models, the proposed $APMM$ looks only for a small subset of documents from the collection D thanks to the study of $selected$ set of documents i.e: those relevant for the nearest neighbors queries to q_x. This will considerably speed up the search process and thus improve the response time.

Algorithm 1: Analogical Proportion-based Matching Model APMM.

```
Input: a set $Q = \{q_r, rel(q_r)\}$, a test query $q_x \notin Q$, a threshold $\theta$, $k \geq 1$
CandidateRelDoc($q_x$)=null, RetrDoc($q_x$)= $\emptyset$
**for** each $q_r \in Q$ **do** compute $CosineSim(q_x, q_r)$ **end for**
sort by increasing order the list $L$ of values $\{CosineSim(q_x, q_r) | q_r \in Q\}$
build up the set $NN_k(q_x) = \{q_r \in Q$ s.t. rank($CosineSim(q_x, q_r)$) $in\ L \leq k\}$
    //Document search
CandidateRelDoc = $\bigcup\{rel(q_r)$ s.t. $q_r \in NN_k(q_x)\}$
    // Document Filter
**for** each document $d \in CandidateRelDoc$ **do**
    Compute $AnalogicalRelevance(q_x, d)$
    **if** $AnalogicalRelevance(q_x, d) < \theta$ **then**
        CandidateRelDoc($q_x$).Remove(d)
    **end if**
**end for**
RetrDoc($q_x$) = CandidateRelDoc($q_x$)
return (RetrDoc($q_x$))
```

5 EXPERIMENTATIONS AND DISCUSSION

In this section, we first provide the detail of the experimental results of the proposed algorithm. Then, we discuss a comparative study between our approach and the most known efficient model Okapi-BM25 (available in the Terrier platform[1]) to show the relevance of the training step in the IR matching model. For this purpose, we conduct a variety of evaluation scenarios of the $APMM$ following the TREC protocol applied to three IR test collections (CRAN, CACM and CISI)[2]. These collections have been selected among other known standards due to the high

[1] http://terrier.org/

[2] http://ir.dcs.gla.ac.uk/resources/test_-collections/

Table 4: The degree of similarity between queries in the three test collections.

CRAN	**MinSim**	0.5	0.54	0.58	0.62	0.66	0.72	0.76	0.78
	#test queries	43	33	24	20	14	12	8	6
CACM	**MinSim**	0.28	0.3	0.32	0.34				
	#test queries	8	7	6	4				
CISI	**MinSim**	0.5	0.54	0.56					
	#test queries	4	3	2					

similarity between their topics. To figure out the degree of similarity between queries in the three test collections, we first compute the cosine similarity between each pair of queries as described in Algorithm 1. Then, queries q_x are grouped into subsets S such that:
$q_x \in S$ *iff* $CosineSimilarity(q_x, NN_1(q_x)) \geq MinSim$, for a given *MinSim* value. Table 4 provides a summary of the number of queries in each subset corresponding to a given *MinSim* value. This table shows that queries from the CRAN test collection reveal a higher degree of similarity between them than for those of CACM and CISI.

For all the following experiments, we run Algorithm 1 with $\theta = 0$ and $k = 5$ to benefit from larger set of nearest neighbor queries. Then, we use the same test queries to run the Okapi model.

5.1 Main Results of the CRAN Test Collection

The first two rows in Figure 1 present a set of recall-precision curves that compare the *APMM* to Okapi for the CRAN test collection for different similarities' scores (*MinSim*) between queries. The second two rows in Figure 1 present the precision values at different top documents (e.g. $P@5$, $P@10$,..., $P@30$), the *MAP* and the *R-Precision*. For instance, the precision at point 30, namely $P@30$, is the ratio of relevant documents among the top 30 retrieved documents.

Figure 1, shows that the *APMM* recall-precision curves outperform the Okapi on the majority points of recall for all similarity levels.

In the second two rows of Figure 1, we can also see that the *APMM* is clearly better than the Okapi in terms of *precision* at top returned documents, the *MAP* and the *R-Precision* for *all* similarity levels. The proposed model is largely better with a significant gap for the first values of precision corresponding to the first selected documents ($P@1$-$P@5$). It is also important to note that, for the CRAN test queries having similarity greater or equal to 72% (*MinSim* = 0.72), the *APMM* achieves the best results in terms of *MAP* and *R-Precision* with the largest gap to Okapi.

5.2 Main Results of the CACM and CISI Test Collections

As noted before (see Table 4), the two test collections CACM and CISI have a reduced number of similar queries if compared to the CRAN. This limits their effectiveness for testing our approach. If we analyze the recall-precision curves provided for the CACM test collection (see: first row in Figure 2), we remark that the *APMM* still outperforms the Okapi on the majority points of recall. However, the Okapi is slightly better between the points of recall 0.5 and 0.8.

For the third test collection CISI (see: second row of Figure 2), we note that the *APMM* is better than the Okapi especially for low-level points of recall (less or equal to 0.4). However, for the high-level points of recall, the reverse is true. The two models achieve similar results on the point of recall 1.

In Figure 2, we also provide a comparison in terms of precision at different *top* documents, *MAP* and *R-Precision* metrics for the two test collections CACM and CISI.

From results of the CACM (see: third row in Figure 2), it is clear that the *APMM* is more efficient, if compared to Okapi, in terms of precision at the first top returned documents from $P@1$ until $P@10$, the *MAP* and the *R-Precision* and for *all* similarity levels.

Regarding the CISI test collection (see: last row in Figure 2), we can draw the following conclusions:

- For test queries having *MinSim* = 0.5 or 0.54, the *APMM* outperforms Okapi in terms of precision at *all* top returned documents (except at $P@5$) and especially the *R-Precision.*
- When the similarity is increased to 56%, the Okapi seems better than *APMM* in terms of precision at top returned documents ($P@4$,..., $P@20$) and the *MAP*, while *APMM* performs clearly better in terms of $P@3$ and *R-Precision.* Both matching models have close results at P@1, $P@2$ and $P@30$.

5.3 Improvement Percentage

In order to investigate more the effectiveness of the *APMM*, we compare its results to the Okapi still using the previous evaluation metrics but in a different way. In Table 5, we assess and present the *best improvement percentages* of the *APMM* if compared to Okapi for the three datasets using the precision at different top documents, the *MAP* and the *R-Precision.* For each dataset, we only present the detail of the improvement at different precision metrics for the sim-

Table 5: The best improvement percentages of *APMM* compared to Okapi for the three test collections.

Dataset (Best MinSim)	*P*@1	*P*@2	*P*@3	*P*@4	*P*@5	*P*@10	*P*@15	*P*@20	*P*@30	*MAP*	*R-precision*
CRAN(Best MinSim=0.72)	200	168	130.38	63.75	67.35	41.45	58.81	46.37	65.59	53.59	73.13
CACM(Best MinSim=0.34)	300	65.33	34	71.43	40	12.94	-8.81	-14.48	1.24	33.81	16.67
CISI(Best MinSim=0.5)	0	24	50	10.22	-9.09	30	25.44	17.14	26.67	-7.8	28.11

Table 6: The p-value for the Wilcoxon matched-pairs signed-ranks test for the three test collections.

CRAN	**MinSim**	0.5	0.54	0.58	0.62	0.66	0.72	0.76	0.78
	p-value	0.003	0.002	0.003	0.003	0.003	0.002	0.002	0.003
CACM	**MinSim**	0.28	0.3	0.32	0.34				
	p-value	0.013	0.015	0.055	0.011				
CISI	**MinSim**	0.5	0.54	0.56					
	p-value	0.009	0.028	0.312					

ilarity level that provides us the *best* improvement percentage. If we analyze the results of the CRAN test collection, we notice an average improvement of *APMM* if compared to Okapi of about 94% (if we consider all the top returned documents $P@1$,..., $P@30$). We also registered an improvement respectively about 54% for the *MAP* and 73% for the *R-Precision*.

Overall, the *APMM* also highlights an improvement if compared to Okapi for the two other test collections CACM and CISI. The average improvement on all the top returned documents is respectively about 56% for the CACM and 20% for the CISI.

These conclusions confirm our first observations presented above about the efficiency of the *APMM* if compared to Okapi especially for the IR collections having large similarity between their test queries as in the case of CRAN.

5.4 Statistical Evaluation of *APMM*

It is important to know if the previously observed improvement of the *APMM* over Okapi is statistically significant. This is can be checked using the Wilcoxon Matched-Pairs Signed-Ranks Test as proposed by Demsar (Demsar, 2006). Table 6 summarizes the results of the computed *p-values* comparing the *APMM* to Okapi in terms of precision at different top documents, the *MAP* and the *R-Precision* scores for respectively the CRAN, CACM and CISI test collections. The null hypothesis (stating that the two compared models perform equally) has to be rejected when the *p*-value is *less* than the threshold 0.05.

The computed *p*-values show that:

Table 7: A comparative study between *APMM* and (Fuhr and Buckley, 1991).

	APMM			(Fuhr and Buckley, 1991)		
	CRAN	**CACM**	**CISI**	**CRAN**	**CACM**	**CISI**
P@15	0.30	**0.38**	**0.46**	**0.37**	0.33	0.17
MAP	**0.46**	**0.30**	**0.22**	0.38	0.29	0.20

- In case of the CRAN test collection (see: Table 6), the improvement of the *APMM* compared to Okapi in terms of precision at different top documents, the MAP and the R-Precision scores, is *statistically* significant for *all* similarity levels between test queries (all $p - values < 0.05$). The best registered $p - value = 0.002 < 0.05$ corresponds to $MinSim = 0.54$ (33 test queries), $MinSim = 0.72$ (12 test queries) and $MinSim = 0.76$ (8 test queries).
- If we consider the CACM test collection, we can see that the improvement of the *APMM* compared to Okapi is still statistically significant in different levels of similarity between test queries ($p - value < 0.05$ in the most cases). Except for similarity score at least equal to 32%, for which we have registered a borderline p-value = 0.05. As can be noted in Figure 2, the Okapi outperforms *APMM* in some precisions at different top documents such as $P@15$, $P@20$ and $P@30$.
- Regarding the CISI test collection and when the similarity between queries is not less than 50%, a statistically significant improvement of *APMM* over Okapi can clearly be observed ($p - value = 0.009$). We can also see a significant improvement when we extend the similarity between the test queries to 54% ($p - value = 0.028 < 0.05$). However if we restrict the minimum similarity to 0.56%, the improvement of the *APMM* over Okapi is not statistically significant since the $p - value = 0.312 > 0.05$. This confirms what we previously noted above.

5.5 Further Comparison and Discussion

In this sub-section, we aim to provide further comparisons of the *APMM* to the state-of-the-art approaches. To the best of our knowledge, there are no previous works applying analogical proportions in the IR context. For this reason, we compare our model to the work proposed by Fuhr and Buckley (Fuhr and Buckley, 1991). First, because the authors have applied a kind of learning in their model and second because they have tested their approach on the same IR Glasgow test collections (CRAN, CACM and CISI) and they have used the same indexing technique (TFxIDF). Table 7 summarizes experimental results of Fuhr and Buckley (Fuhr and Buckley, 1991) and *APMM* for the three test collections. These results show an improvement of *APMM* if compared to Fuhr and Buckley (Fuhr and Buckley, 1991) model except in P@15 for the CRAN dataset for which their approach outperformed the APMM (see bold values in Table 7).

6 CONCLUSION

The success of analogical proportions in a variety of domains, such as in classification and language processing, led us to wonder whether it may be a successful tool for building an IR matching model. We are mainly interested to this last field in this paper. We have first studied the way to apply analogy between queries and documents. Then, given a particular indexing query term, we formalize two logical proportions linking queries and their corresponding relevant documents for an analogical inference. These proportions form the basis for our matching model.

The two proposed analogical proportions help to define *agreement* and *disagreement* scores useful to estimate to what extent any document, from the collection, is to be accepted or rejected given a new query. The agreement score is calculated according to the common terms between the query and the document while the disagreement is computed using the number of terms they differ. The two scores treat documents differently: the disagreement allows you to exclude irrelevant documents from the returned list, while the agreement score strengthen the relevance of the remaining documents not eliminated by the disagreement. Based on these two scores, we have proposed and tested a new analogy-based IR matching model on three IR Glasgow test collections. The experimental results highlighted the effectiveness of the model compared to the well known efficient Okapi IR model.

The analogy-based IR matching model can be applied in different IR/CLIR tasks such as in query expansion, disambiguation and translation tasks that will be our future interest.

REFERENCES

Amit, S. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24:35–43.

Bayoudh, S., Miclet, L., and Delhay, A. (2007). Learning by analogy: A classification rule for binary and nominal data. In *Proc. IJCAI 2007*, pages 678–683.

Bounhas, M., Prade, H., and Richard, G. (2014). Analogical classification: A new way to deal with examples. In *Proc. ECAI*, pages 135–140.

Bounhas, M., Prade, H., and Richard, G. (2017a). Analogy-based classifiers for nominal or numerical data. *IJAR*, 91:36–55.

Bounhas, M., Prade, H., and Richard, G. (2017b). Oddness/evenness-based classifiers for boolean or numerical data. *IJAR*, 82:81–100.

Bounhas, M., Prade, H., and Richard, G. (2018). Oddness-based classification: A new way of exploiting neighbors. *IJIS*, 33(12):2379–2401.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Fuhr, N. and Buckley, C. (1991). A probabilistic learning approach for document indexing. *ACM Trans. on Inf. Sys.*, 9(3):223–248.

Hesse, M. (1959). On defining analogy. *Proceedings of the Aristotelian Society*, 60:79–100.

Lepage, Y. (2001). Analogy and formal languages. In *Proc. FG/MOL 2001*, pages 373–378.

Miclet, L., Bayoudh, S., and Delhay, A. (2008). Analogical dissimilarity: Definition, algorithms and two experiments in machine learning. *Journal of Artificial Intelligence Research*, 32:793–824.

Miclet, L. and Prade, H. (2009). Handling analogical proportions in classical logic and fuzzy logics settings. In *Proc. ECSQARU'09*, pages 638–650. Springer, LNCS 5590.

Prade, H. and Richard, G. (2010). Reasoning with logical proportions. In *Proc. KR 2010*, pages 545–555.

Prade, H. and Richard, G. (2012). Homogeneous logical proportions: Their uniqueness and their role in similarity-based prediction. In *Proc. KR 2012*, pages 402–412.

Prade, H. and Richard, G. (2013). From analogical proportion to logical proportions. *Logica Universalis*, 7(4):441–505.

Prade, H., Richard, G., and Yao, B. (2012). Enforcing regularity by means of analogy-related proportions-a new approach to classification. *Int. J. of Comp. Inf. Sys. and Indus. Management App.*, 4:648–658.

Stroppa, N. and Yvon, F. (2005a). An analogical learner for morphological analysis. In *Proc. CoNLL 2005*, pages 120–127.

Stroppa, N. and Yvon, F. (2005b). Analogical learning and formal proportions: Definitions and methodological issues. Technical report.

Yvon, F., Stroppa, N., Delhay, A., and Miclet, L. (2004). Solving analogical equations on words. Technical report, Ecole Nationale Supérieure des Télécommunications.

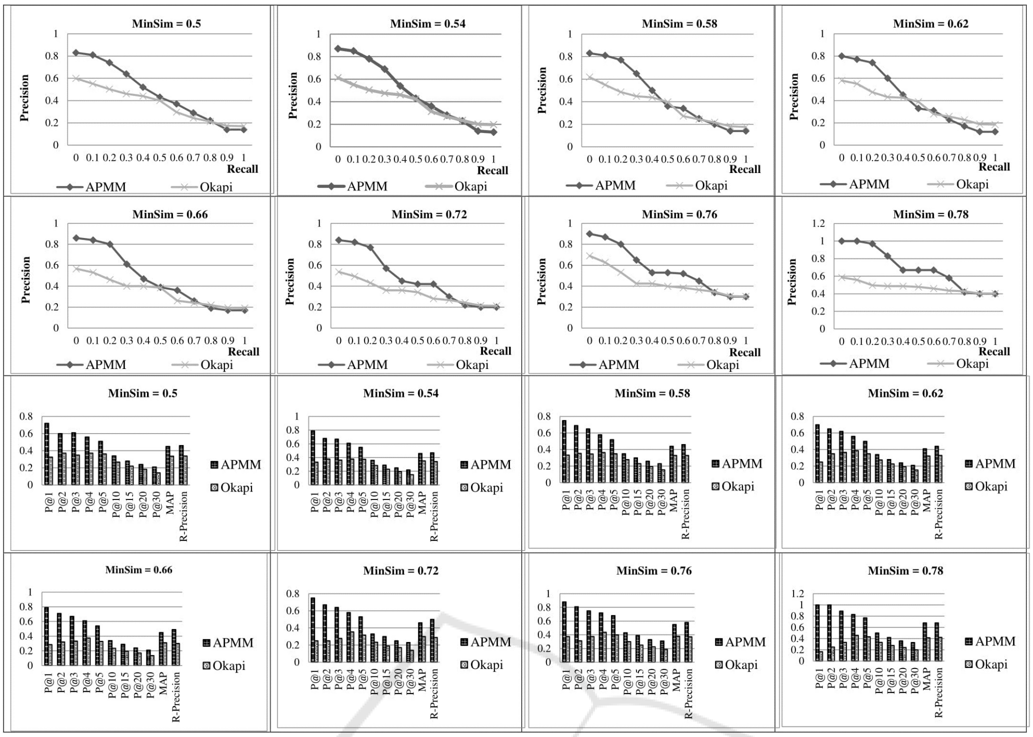

Figure 1: Main results of the CRAN test collection: APMM vs. Okapi.

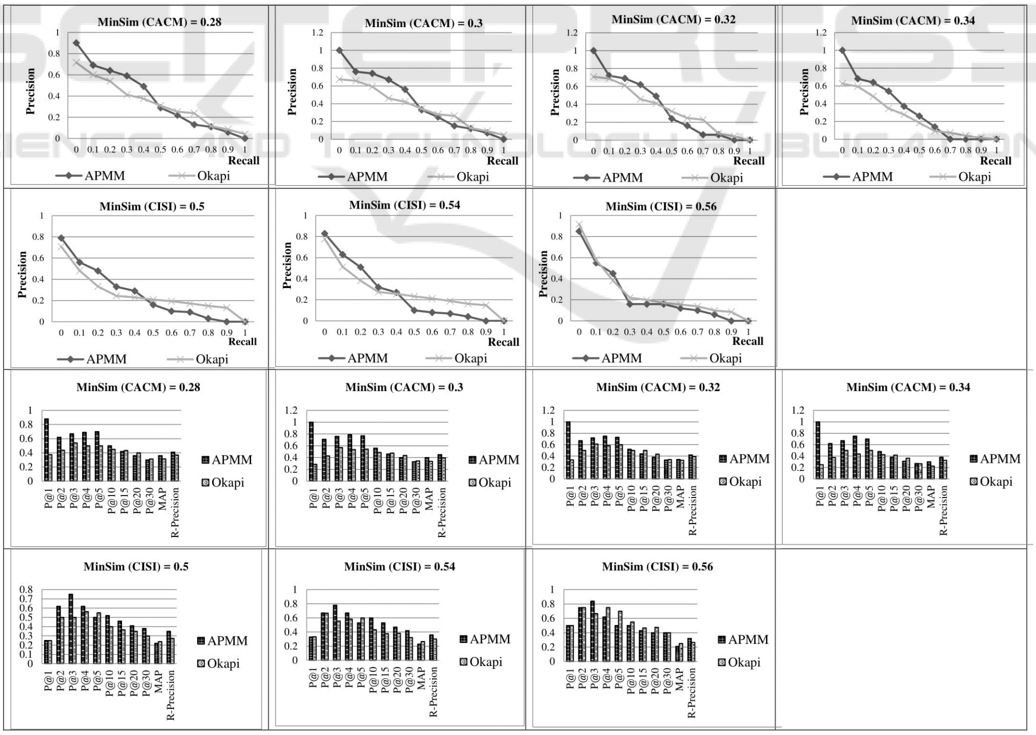

Figure 2: Main results of the CACM and CISI test collection: APMM vs. Okapi.