Miracl at Clef 2014 : Ehealth Information Retrieval Task

Nesrine KSENTINI, Mohamed Tmar, and Faïez GARGOURI

MIRACL Laboratory, City ons Sfax, University of Sfax, B.P.3023 Sfax TUNISIA
ksentini.nesrine@ieee.org,
mohamed.tmar@isimsf.rnu.tn,
faiez.gargouri@isimsf.rnu.tn,

Abstract. This paper presents our first participation in user-centred health information retrieval task at the CLEFeHealth 2014. This task has as objective the information retrieval to answer patients' questions when reading clinical reports. We have submitted only the mandatory run (baseline system).
The obtained results are motivating with map=0.1677 and p@10=0.5460 but can be improved.

Keywords: information retrieval, vector space model, medical documents.

1 Introduction

With the increasing number of documents on the web, searching relevant documents to users queries becomes a difficult task especially if the queries are short and do not represent well the user need. The process of searching documents in a corpus in order to meet a need is called Information Retrieval (IR).
This process (IR) focuses on the representation, storage, organization of information that should allow the user quick and easy access to information.
To automate the task of RI, Information Retrieval Systems (IRS) are developed to provide all necessary functions for information retrieval illustrates in the figure 1.

The system builds an index of the documents which is an essential data structure because it allows fast searching over large volumes of data.
Given that the document database is indexed, the searching process can be started. The user specifies a query which is parsed and transformed by the same operations applied to the document database. After that the system retrieves documents that are relevant to the query from the index and displays them to the user with a ranked way.
The user can examine the ranked documents. He can identify a subset of the documents seen as relevant and initiate a relevance feedback step. The system uses the documents selected by the user in order to refine the query.
The evaluation of such systems appears to be a necessity. This evaluation is

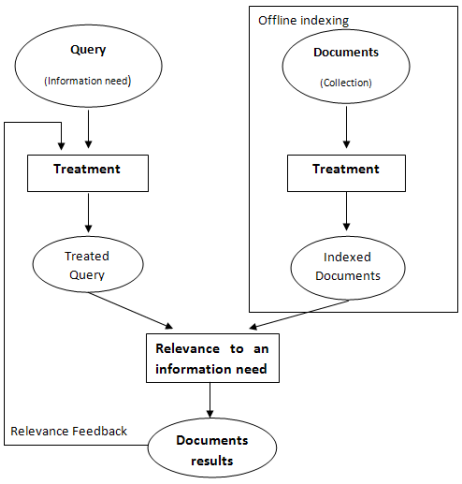

Fig. 1. Process of Information Retrieval System

based on the notion of relevance [1]. To improve the relevance of IR in IRS, several studies have been proposed several IR models: the Boolean model, the probabilistic model, the vector model.

– **Boolean Model:**

  It is a model which is based on the set theory. Thus, the user query is represented as a logical equation composed of keywords and Boolean (logic) operators (OR, AND, NOT) and documents are represented by a list of keywords. The research process with this model performs operations of union, intersection and difference, defined by the existence or absence of index terms, to achieve an exact match between the documents and the equation of the query.

– **Probabilistic model:**

  The probabilistic model addresses the problem of information retrieval in a probabilistic framework. It allows modeling of the selection process documents in an IRS based on probability theory. The basic principle of the probabilistic model is to present the search results in an IRS order (O) based on the probability of relevance of a document with respect to a query. The basic idea is first to calculate two conditional probabilities $P(R/D)$ and $P(NR/D)$ with a given query R is the relevance (all relevant documents) and NR irrelevance (all irrelevant documents). The terms are not weighted, but only take the value 0 (term absent) or 1 (this term). The principle of the model is to calculate an ordering function ($O(D)$) that would classify documents that meet a query with

$$O(D) = \frac{P(R/D)}{P(NR/R)} \tag{1}$$

– **Vector Space Model:**

  The vector space model (VSM) was developed by Salton for the SMART information retrieval system [2]. It is an algebraic model that uses a vector representation for documents as well as queries in a large dimensional space. The performance of this model depends on the weighting of terms in these vectors which represents the degree of relationship between a term and a document. The relevance of a document relative to a query is defined by distance measurements in a vector space. The most widely used numeric similarity measure is the cosine of the angle between the vector and the vectors [2, 4, 3] .

In our participation, we used Vector Space Model because it use the weighted vector not the binary one and it allows computing a degree of similarity between documents and queries; also it returns ranking documents according to their relevance.

2 Material and Methods

2.1 Database

Document Collection: The data set for the ehealth information retrieval task consists of a set of medical-related documents, provided by the Khresmoi project [13, 12, 8]. This collection contains documents covering a broad set of medical topics, and does not contain any patient information. The documents in the collection come from several online sources, including the Health On the Net organisation certified websites, as well as well-known medical sites and databases (e.g. Genetics Home Reference, ClinicalTrial.gov, Diagnosia).

Topics: The test set consists of 50 medical professional queries containing the following fields [13, 8]:

– Title: text of the query.
– Description: longer description of what the query means.
– Narrative: expected content of the relevant documents.
– Profile: main information on the patient (age, gender, condition).
– Discharge summary: ID of the matching discharge summary provided by Ehealth Information Extraction task.

2.2 Methods

Our work was to index and search the top-1000 relevant medical documents for each topic using a plain-text search engine.
We chose to use the terrier platform in our case [6, 7]. It is an efficient, effective and flexible open source search engine, easily deployable on large-scale collections of documents [5]. Terrier implements state-of-the-art of indexing and retrieval functionalities. It is an open source, and a comprehensive and transparent platform for research and experimentation in text retrieval. Terrier is written in Java, and is developed at the School of Computing Science, University of Glasgow.

The first step made before you begin indexing was the extraction of individual HTML documents from raw files. Each of individual document was saved in its own file having the name of its UID mentioned in the raw file.

After extracting medical documents, we started the indexing step. It happened offline and contains four steps:

Tokenization: is a process of breaking a text up into words called tokens.
Stop wording: most common words that would appear in the list of tokens are excluded because they have a little value when searching documents to a user need.
Stemming: is a process of linguistic normalization for reducing variant forms of tokens to their stem or root form.
Storing: information (terms) were stored in file with special structure called inverted file by specifying the number of occurrences (term frequency (TF) and their location. This file allows rapid access during query time.

The third step is searching relevant documents for a particular query; it was carried online. Firstly, we processed query in the same way documents were processed during indexing. Then we represented documents and query with weighted vectors of terms which represents the degree of relationship between a term and a document in the whole collection. This degree is obtained by multiplying the two measures (TF) and (IDF) with

$$IDF_{term_i} = \log \frac{\mid D \mid}{\mid d_j : term_i \in d_j \mid} \tag{2}$$

Where
$\mid D \mid$: total number of documents in the collection.
$\mid d_j : term_i \in d_j \mid$: number of documents where the $term_i$ appears

Then we calculated similarity (scores) between vectors using cosine similarity measure and ranked the documents according to their scores in descending order. The top 1000 documents with the highest scores were the best match to the query and returned as relevant documents.

3 Results

Table 1. Obtained results for all queries

Type	Relevant	Retrieved	MAP	p@10
Our run	3209	1189	0.1677	0.5460

In table 1 we show results obtained by taking account of pooling all search results.
Our system can retrive 38% (1189 among 3209) of relevant documents and has a measure of map equal to 0.1677 and of p@10 equal to 0.5460.
These results are motivating but can be improved by including for example the notion of semantics in the search process [11, 10, 9].

The plots below in figure 2 compares our run against the median and best performance (p@10) across all systems submitted to Ehealth task for each query.
In particular, for each query, the height of a bar represents the gain/loss of your system and the best system (for that query) over the median system. The height of a bar in then given by:

$$greybars : height(q) = our_p@10(q) - median_p@10(q)$$

$$whitebars : height(q) = best_p@10(q) - median_p@10(q)$$

We can see that our run has 3 queries that achieve the best performance, 17 queries perform better than the median, while 26 queries were worse than the median, and other 5 queries perform in the median line. It means that in a specific task such as medical document retrieval, results are motivating but need improvement.
In terms of the best performance of each query, some queries can achieve a very high result like query 14,22,32,33,39,48 and 49 (more than 70%), while some queries perform near the median like query 2 and 20 (only about 10%).

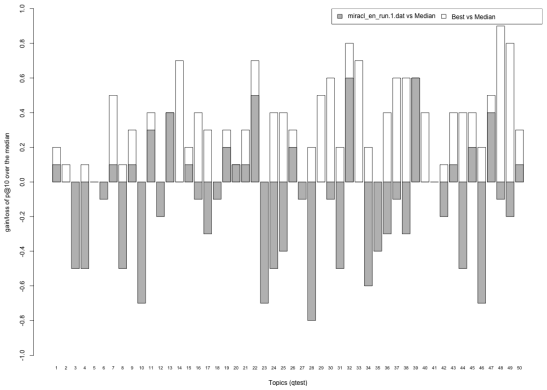

Fig. 2. Comparison between our Run and the other systems against the median and best performance

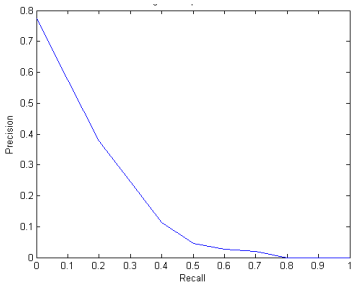

Fig. 3. Average Recall-precision curve

We show in figure 3, the average recall-precision curve for all 50 topics. There is a tradeoff between recall and precision. We notice that when increasing recall by retrieving more, precision decreases.

4 Conclusion and future work

In our first participation in ehealth information retrieval task at the CLEFe-Health 2014, we obtained motivating results for the search in a large collection of medical documents to answer patients' questions when reading clinical reports. For future work, we will try to improve the obtained results by including the notion of semantics between terms of documents [11].

References

1. Wei, Xing and Croft, W Bruce: Investigating retrieval performance with manually-built topic models. pp.333–349 (2007)
2. Salton, Gerard and Wong, Anita and Yang, Chung-Shu: A vector space model for automatic indexing. In :Journal of Communications of the ACM. pp.613–620, vol.18 (1975)
3. Aswani Kumar, Ch and Radvansky, M and Annapurna, J: Analysis of a Vector Space Model, Latent Semantic Indexing and Formal Concept Analysis for Information Retrieval. In :Journal of Cybernetics and Information Technologies. pp.34–48, vol.12 (2012)
4. Turney, Peter D and Pantel, Patrick and others: From frequency to meaning: Vector space models of semantics. In :Journal of artificial intelligence research. pp.141–188, vol.37 (2010)
5. Santos, Rodrygo LT and McCreadie, Richard and Plachouras, Vassilis: Large-scale information retrieval experimentation with Terrier. pp.2601–2602 (2011)
6. Ounis, I. and Amati, G. and Plachouras, V. and He, B. and Macdonald, C. and Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. (2006)
7. Ounis,I.;Lioma,C.;Macdonald,C.;Plachouras,V.: Research Directions in Terrier. In :Journal of Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper. (2007)
8. Liadh Kelly and Lorraine Goeuriot and Hanna Suominen and Danielle L. Mowery and Sumithra Velupillai and Wendy W. Chapman and Guido Zuccon and Joao Palotti: Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: Proceedings of CLEF 2014 (2014)
9. Agirre, Eneko and Alfonseca, Enrique and Hall, Keith and Kravalova, Jana and Paşca, Marius and Soroa, Aitor: A study on similarity and relatedness using distributional and WordNet-based approaches. In: Proceedings of Human Language Technologies. pp.19–27 (2009)
10. Eneko Agirre and Montse Cuadros and German Rigau and Aitor Soroa: Exploring Knowledge Bases for Similarity. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation. (2010)
11. Ksentini Nesrine and Tmar Mohamed and Gargouri Faiez: Detection of Semantic Relationships between Terms with a New Statistical Method. In: Proceedings of the 10th International Conference on Web Information Systems and Technologies (WEBIST 2014). pp.340–343 (2014)
12. Goeuriot, Lorraine and Hanbury, Allan and Jones, Gareth JF and Kelly, Liadh and Kriewel, Sascha and Martinez Rodriguez, Ivan and Muller, Henning and Tinte, Miguel: Supporting collaborative improvement of resources in the Khresmoi health information system. (2012)
13. Lorraine Goeuriot and Liadh Kelly and Wei Li and Joao Palotti and Pavel Pecina and Guido Zuccon and Allan Hanbury and Gareth Jones and Henning Mueller: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In: Proceedings of CLEF 2014 (2014)