Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 139–145

Procedia Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

Support Vector Machines for a new Hybrid Information Retrieval System

Hamid KHALIFI[a],*, Abderrahim ELQADI[b], Youssef GHANOU[a]

[a]TIM team, EST Meknes - Moulay Ismail University, Meknes
h.khalifi@gmail.com
youssefghanou@yahoo.fr
[b]LASTIMI, EST Sale - Mohammed V University, Rabat
elqadi_a@yahoo.com

Abstract

Information Retrieval systems are used to extract, from a large database, relevant information for users. When the type of data is text, the complex nature of the database makes the process of retrieving information more difficult. Generally, such processes reformulate queries according to associations among information items before the query session. In this latter, semantic relationships or other approaches such as machine learning techniques can be applied to select the appropriate results to return. This paper presents a formal model and a new search algorithm. The proposed algorithm is applied to find associations between information items, and then use them to structure search results. It incorporates a natural language preprocessing stage, a statistical representation of short documents and queries and a machine learning model to select relevant results. On a series of experiments through Yahoo dataset, the proposed hybrid information retrieval system returned significantly satisfying results.

*Keywords:*Information retrieval; natural language processing; unsupervised classification; supervised classification; support vector machines.

1. Introduction

Nowadays, the word is witnessing a steady increase in the volume of available data[1]. Data are stored in diverse sources and every day, users consult these sources searching various kinds of information. This had led to the emergency of robust information retrieval systems to deal with this massive amount of unstructured data.

* Corresponding author. Tel.: "+212 6 68 61 55 50" ;
E-mail address: h.khalifi@gmail.com

10.1016/j.procs.2018.01.108

To explore this profusion, some modern search engines resort to semantics analysis and lexical matching. This is due to the fact that a same concept can be expressed by using different vocabularies and language styles [2].

In parallel, text classification is a subdomain of Natural Language Processing which correlates with information retrieval. Recently, models based on machine learning have become increasingly popular [3]. While these models achieve very good performance in practice, they tend to be relatively slow, limiting their use on very large datasets [4]. Moreover, they achieve weakest performance when the work introduces semantics.

Facing this dilemma between categorization performance on the first hand and semantic analysis on the second hand, lexical matching and machine learning can be gathered into a same hybrid model in order to meet the growing information retrieval needs.

Information Retrieval techniques are not limited to unstructured text data. They also have been applied to databases of images, stored speech, and other forms of data. They are more difficult to apply when the retrieved objects introduce semantics.

Generally, an information retrieval process consists of four main steps [5]: indexing, query formulation, comparison and finally the feedback.

Language modeling approaches to information retrieval connect the problem of retrieval with that of language model estimation. The main idea of such model is to estimate a language model for each document, and then rank documents by the likelihood of the query according to the estimated language model. The core problem in language modeling is the inaccuracy due to data sparseness [6].

Unfortunately, many typical challenges are noticed in language modeling approaches especially the process of stemming for unstructured text data. Therefore, this work will primarily focus on resolving the vocabulary mismatch problem. Variant word forms will be mapped to their base form (stemming).

In this sense, stemming is a crucial pre-processing step in language modeling as well as a very common requirement of Natural Language processing applications[7].

This paper is organized as follows. Section 2 summarizes main works in the area of information retrieval, and highlights the techniques of machine learning. Section 3 describes different steps of the proposed information retrieval system. Section 4 discusses experimental results. These results indicate the outperformance of the sophisticated proposed algorithm. This section includes also an empirical evaluation, as well as a discussion of its different steps. Finally, section 5 concludes the paper with a brief presentation of our further research topics.

2. Background

Information Retrieval systems have been widely investigated during last decades. Some approaches focused on lexical and vocabulary analysis while other ones based on Machine Learning techniques.

2.1. Information Retrieval techniques

Probabilistic indexing models have been considered by a number of studies as standard models for document retrieval. Such models assume that a subset of terms occurring in the document would be significant for indexing and the documents should be approximately equal length [8]. These models attempt to explicitly compute the probability of relevance between a document and the query [9]. The proportion of relevant documents is not enough to make the approximation for the initial retrieval [5].

Most information retrieval models in the past have based their statistical independences on strong or rarely satisfied assumptions: this explains the limited performance they achieved.

The critical language issue for previous works is the term mismatch problem [10]: users do often not use the same words to refer to objects or ideas. This is compounded by synonymy and polysemy [11].

The second class of approaches uses machine learning; different techniques have also been widely used in information retrieval systems which rank documents basing on statistical computations [5]. The user query is then considered as an ideal relevant document, and a similarity measure is computed between this user query and the rest of documents.

Le and Mikolov [12] proposed an unsupervised learning method to learn a paragraph vector as a distributed representation of sentences and documents.

Similarly, an unsupervised sentence embedding method was proposed later by Kiros, et al [13]. This method yielded good results on large text corpus. The main idea of the model was to link each sentence with its previous and next sentences.

Afterwards, another approach for sentence embedding based on deep learning was proposed by Palangi, et al. [14]. Instead of keeping only the largest entries among word vectors in one vector, authors kept also the k largest entries in k different vectors. This model has shown good performance in sentiment prediction and question type classification tasks.

In summary, there are two main classes of techniques for information retrieval. The first class treats the individual queries as documents and extract features the original user query [15], while the second technique analyzes the relation between queries and the retrieval results to enlarge the context of retrieving.

2.2. Machine Learning

Machine Learning techniques have been extensively applied to text mining fields in general and to information retrieval in particular.

This work compared main algorithms and retained k-means as a model for unsupervised classification and Support Vector Machines as a model for supervised classification.

The principle of k-means algorithm consists of partitioning n observations into k classes ($k \leq n$) by minimizing the distance between the observations inside each class. The algorithm starts by designating k centers of classes, and then it iteratively carries out two steps:

- Define, for each observation that is not a class center, the closest class center.
- Define the center of each new class

The k-means algorithm is one of the most used unsupervised classification algorithms [16].

Briefly, the basic idea of Support Vector Machine is as follows [17]:

Let's consider a set of data: $(x_1, y_1) \ldots (x_k, y_k)$ where $x_i \in R^n$ are the input data and $y_i \in \{-1,1\}$ their class labels.

The aim is to construct a hyperplane $\omega^T x + b = 0$, that separates two classes defined by $\omega^T x + b = 1$ and $\omega^T x + b = -1$ [18].

The problem consists on minimizing $\frac{1}{2}\omega^T\omega$.

The formulation is:

$$\begin{cases} \min \frac{1}{2}\omega^T \omega \\ s.t: \\ y_i((\omega^T x_i) + b) \geq 1 \end{cases}$$

By introducing ξ_i and ξ_i^* (upper and lower constraints on the outputs) and C a constant that regularizes the equation, the problem becomes:

$$\min \frac{1}{2}\omega^T \omega + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

The mapping $x_i \in R^n$ ($1 \leq i \leq n$) is performed by a kernel function:

$$K(x_i, x_j) = \phi(x_i)\phi(x_j)$$

Different kernel functions are compared in this study: linear [19], radial [20] and polynomial [21].

The decision function implemented by SVM can be written as:

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b)$$

The coefficients α_i are the solutions of:

$$\begin{cases} \max \sum_{i=1}^{k} \alpha_i - \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j y_i y_j \left[\phi(x_i) \phi(x_j) \right] \\ s.t: \\ 0 \le \alpha_i \le C \qquad \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}$$

In the next section, the choice of K-Means and SVM will be justified as well as the other steps of the global approach.

3. The proposed approach

The global proposed approach is summarized in the following figure:

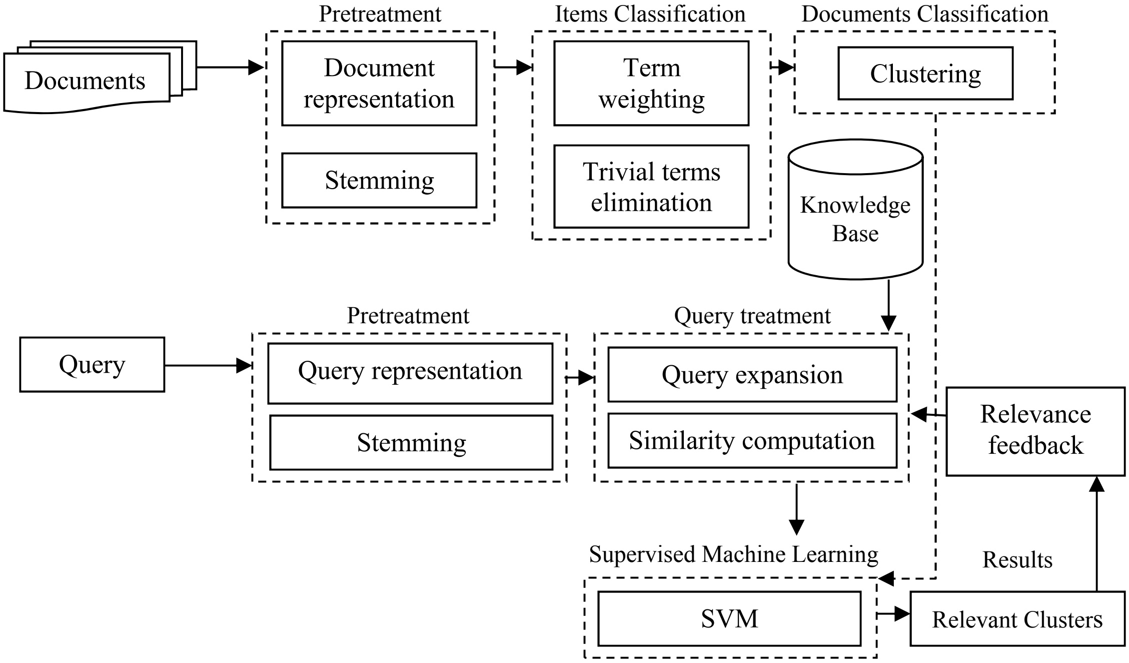

Figure 1: The Hybrid Information Retrieval system

The first step in figure 1 is text pre-treatment: documents sentences are stemmed in order to gather terms referring to the same concept (this eliminates the mismatch problem). Each document is then represented as a vector of occurrences of the stems it contains.

Once the documents represented by vectors in the first step, the second step consists on classifying terms into three categories:

- Trivial terms, such as stop words, which are eliminated in order to reduce the ambiguity of classification.
- Decisive terms, which belong to a specific class, and can then be considered as an indicator of this class.
- Standard terms, are the terms that are less common than trivial terms and less decisive than decisive terms.

The two last classes of terms are defined after the clustering stage, in which documents are classified into classes of similar vectors according to the retained terms and the similarity used. This process is performed offline, and it allows the classification of new queries online more accurately.

Support vector machines are then used to construct an offline model which associates each representing vector with its formed cluster.

Finally, a given user query is first stemmed and represented by a representing vector as for documents. And its membership is determined by the support vector machines decision function. This membership allows the selection of documents from the returned class to return first.

4. Results and Analysis

To evaluate the performance of the proposed hybrid approach, a dataset from the corpus Yahoo which consists of 22 938 unlabeled documents is used [22], and subset of short query document is retained for test.
This corpus has derived 519 596 terms and marks.

4.1. Computation

The overall performance was evaluated in terms of f-measure which includes implicitly two other measures: recall and precision [23]:

$$precision = \frac{a}{a+b} \quad ; \quad recall = \frac{a}{a+c}; \quad F-measure = \frac{2\times precision \times recall}{precision + recall}$$

Where a is the number of documents that are correctly classified, b: the number of false positives and c is the number of false negatives.

4.2. Results

By limiting the evaluation of the proposed approach to the classification of documents, the results translate its effectiveness on documents categorization.

Table 1. Comparison of SVM kernels while classifying documents

SVM kernel	linear	polynomial	Radial
Precision	98.52%	97.58%	100%
Recall	97.74%	97.94%	100%
F-measure	98.12%	97.75%	100%

One of the main purposes of a classification model in general, and the proposed study in particular is the determination of the steps influencing the performance of the global process [24]. The evaluation of their relevance proved quite interesting and useful. The estimations of their relative contribution are as follows:

Table 2. Contribution of different steps in the overall performance

Step	Stemming	Terms classification	Kernel SVM
Approach with	100%	100%	100%
Approach without	99.02%	96.73%	97.75%
Improvement	0.98%	3.27%	2.25%
Contribution	15.08%	50.31%	34.62%

By combining the four steps of the global approach, documents are successfully classified into formed clusters by respecting the influences of decisive terms. A given new query is then treated as a new document and most relevant documents (those belonging to the same class as user query) are returned first.

By augmenting the value of k (in k-means model), the clusters become smaller and the retrieving process becomes faster. However, forming cluster of few instances may lead to the problem of over learning [25] and then reduce the global performance.

Moreover, analyzing users' feedbacks allows the evaluation of the proposed model from another perspective. The relevance of each query will be validated according to user's response. The implementation of such online systems is part of our current research.

5. Conclusion and Perspectives

Information retrieval is a long-studied topic with a wide range of applications. Existing models are either lexical or machine learning (probabilistic and statistical). However, each type of previous models has its limits and no model has an absolute precision and performance.

In recent years big efforts have been devoted to improve the performance of information retrieval systems and many different directions have been explored for this purpose. In this sense, this paper introduced a new hybrid approach for information retrieval. Four steps have been integrated into the proposed approach: stemming, representation by vectors, terms classification and documents classification. This latter combined k-means and Support vector machines to select relevant documents to return. K-means gathered similar documents while Support vector machines build up a model to predict the class of a given query and then returns documents from the same class.

Finally, our future works will be directed at identifying the opinion of the user in order to further filter results and to focus on documents that would meet the need of the user.

References

[1] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., ... & Twigger, S. (2008). Big data: The future of biocuration. Nature, 455(7209), 47-50.

[2] Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014, November). A latent semantic model with convolutional-pooling structure for information retrieval. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (pp. 101-110). ACM.

[3] Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2016). Very deep convolutional networks for natural language processing. arXiv preprint arXiv:1606.01781.

[4] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

[5] Lewis, D. D. (2014, June). Learning in intelligent information retrieval. In Machine Learning: Proceedings of the Eighth International Workshop (pp. 235-239).

[6] Zhai, C., & Lafferty, J. (2017, August). A study of smoothing methods for language models applied to ad hoc information retrieval. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 268-276). ACM.

[7] Singh, J., & Gupta, V. (2016). Text Stemming: Approaches, Applications, and Challenges. ACM Computing Surveys (CSUR), 49(3), 45.

[8] Ponte, J. M., & Croft, W. B. (2017, August). A Language Modeling Approach to Information Retrieval. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 202-208). ACM.

[9] Harper, D. J., & Van Rijsbergen, C. J. (1978). An evaluation of feedback in document retrieval using co-occurrence data. Journal of documentation, 34(3), 189-216.

[10] Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR), 44(1), 1.

[11] Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 964-971.

[12] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

[13] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294-3302).

[14] Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., ... & Ward, R. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 24(4), 694-707.

[15] Yin, Z., Shokouhi, M., & Craswell, N. (2009, April). Query Expansion Using External Evidence. In ECIR (Vol. 9, pp. 362-374).

[16] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

[17] Chang, C. C., & Lin, C. J. (2006). Training v-support vector classifiers: theory and algorithms. Training, 13(9).

[18] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 27.

[19] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

[20] Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.

[21] Smits, G. F., & Jordaan, E. M. (2002). Improved SVM regression using mixtures of kernels. In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on (Vol. 3, pp. 2785-2790). IEEE.

[22] Yahoo! Webscope dataset ydata-ymusic-user-artist-ratings-v1_0

[http://research.yahoo.com/Academic_Relations].

[23] Powers, D.M.W., 2011. Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

[24] Cherif, W., Madani, A., & Kissi, M. (2015). Towards an efficient opinion measurement in Arabic comments. Procedia Computer Science, 73, 122-129.

[25] Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge.

1. Utilizzo il clustering per raggruppare documenti simili assieme
2. Utilizzo una SVM per classificare (classificazione multiclasse?) una query, ovvero cerco qual è il clustering piu adatto per la query.

DUBBIO: quindi una SVM per ciascun cluster?

3. Una volta assegnato il cluster, ritorno i documenti più importanti per quello.