A new Approach of Documents Indexing Using subject modelling and Summarization

L. OUADIF1, R. EL AYACHI1 and M.BINIZ2

1 Laboratory of Information Processing and Decision Support (TIAD), Faculty of Sciences and Technics, Sultan Moulay Slimane University, Beni Mellal, Morocco.

2 Laboratory LIMATI, Pluridisciplinary Faculty Beni Mellal, Sultan Moulay Slimane University, Beni Mellal, Morocco.

Abstract. Document indexing is a field of research in Natural Language Processing (NLP) that has been rapidly evolving for 70 years. It is an operation that focuses on the synthetic representation of a document according to a model in order to facilitate their subsequent use. This work is concerned with document indexing. Two points are addressed. This work is concerned with document indexing, we are trying to accelerate the indexing process of large document datasets, two points are addressed. The first one concerns the development of a document indexing system using the system's operating process based on three phases namely pre-processing, weighting, and subject modelling. The second point concerns the proposal for a new system that integrates a new developed automatic summary subsystem, the goal of this point is to minimize indexing time.

1. Introduction

The work of a documentalist is not just about classify documents or search for information, it also has to index and summarizes a large number of documents in order to find them during a document search and to provide a quick visualization tool. The task is even more important when it comes to a large documentary unit, but the technology has allowed a lot of progress especially since the entry of computers in this field.

The explosion of information technology has been an important phenomenon in recent years, and the automatic processing of documentary information is an application in constant development. As time goes on, the implementation of computer equipment in documentation centres and libraries continues to grow and continue. This processing is presented by the automation of some tasks such as file management, information retrieval.

This processing is presented by the automation of some tasks such as file management, information retrieval.

In the literature, there are a lot of work focused on documents indexing, likes:

- [8]: The authors presented an adaptation of the KX system that accomplishes unsupervised, multilingual keyword extraction. KX can manage texts in several languages (English, Italian, French, Finnish and Swedish) and is distributed with TsextPro NLP Suite1 [9]

- [10]: This work presents a hybrid and multi-modular system for extracting keywords from the corpus of scientific articles. It is based on semantic spaces constructed using Reflective Random Indexing (RRI) [11]. the system is improved by enriching the semantic spaces with information from a
surface linguistic analysis, and by defining a decision procedure based on a combination of Bayesian networks and rule-based systems.

- [12]: This work presents a contribution on the task of indexing bibliographic records by keywords. The proposed approach combines two parts, one graphic and the other semantics. The first searches in the document's instructions for words graphically close to the keywords used in the training corpora. The second assigns to a new document the keywords associated with the documents of the training corpus which are semantically closest to it.

- [13]: The proposed approach is based on that of TopicRank [14]. It first selects candidate keywords within the document to be analyzed, it groups them into subjects, projects the subjects into a graph and orders them in the manner of Text Rank.

The rest of the chapter is organized as follows, Section 2 present a document indexing system using different approaches of subject modeling, Section 3 describes present our proposed approach to generate document summaries based on the LSA representation, It is an Extractive approach. In section 3 we present a result of the hybridization of the two developed systems, to improve the indexing process by indexing the summaries of documents instead of the complete document.

2. Indexing system using subject modelling

We have realized a basic information retrieval system, that allows to find documents that meets the need of user expressed in the form of a query, our system goes through two essential steps that can be summarized as follows:

Representation of documents and requests (indexing): The representation of documents is realized when the corpus is imported and is updated when a new document is added, and concerning the query, it is indexed when the user formulates his query.

Query-document matching: This task consists of mapping the query to the document by measuring the degree of membership. This amounts to calculating the degree of similarity between the representation of the documents and the representation of the queries.

![Figure 1. architecture of the developed system](image)

2.1. Indexing process

The indexation process used goes through several steps (Figure 2):
The preprocessing phase: we have integrated several techniques to clean the document: Segmentation or tokenization, elimination of empty words, rooting and lemmatization.

The weighting phase consists of assigning weight to each word in the document, using the TF-IDF measure and comparing it with the results of an unweighted system that simply uses the BOW model.

The modelling phase: A subject model can be defined as an unsupervised technique for discovering topics in various text documents. These subjects are abstract in nature, words related to each other form a subject. Similarly, there may be several topics in an individual document [1]. The data were represented using the subject's modelling approaches LSA [6] and LDA [7], and compare the results by a representation based on word embedding (representation by the doc2vec approach).

![Figure 2. Our indexing process](image)

2.2. Query / document matching process
Once the corpus is indexed, the query analysed and represented in the same space as the documents, then we used the cosine measurement as a method of comparison between the two representations and those, in order to determine their degrees of correspondence and to rank the relevant documents in descending order.

It is a matter of calculating the cosine of the angle between the vector representations of the document and the query. The similarity obtained

\[ \text{Sim}_{\text{cosine}}(D, Q) = \frac{\sum_{i=1}^{n} q_i d_i}{||D|| ||Q||} = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2 \cdot \sum_{i=1}^{n} d_i^2}} \]

The classified documents constitute the result of the search, the latter will be returned to the user in the form of a list, we have only taken documents with a degree of correspondence greater than 0.6.

2.3. Experiment and Results
We used three standard datasets:

- A Corpus of Plagiarized Short Answers [2]: the corpus contains 100 documents of 5 topics ("Inheritance", "PageRank", "Vector Space Model", "Bayes Theorem", "dynamic programming"), the average length of documents are 300 terms. Five test queries are given with relevance judgments.

- NPL (also known as VASWANI) is a collection of 11429 document titles. With 93 queries and assessments of relevancy of queries. The average length of documents is 21.6 terms and the average length of queries is 6.6 terms.

- The Cranfield collection. This was the first collection of tests to accurately measure the effectiveness of information retrieval. Collected in the UK from the late 1950s, contains 1400 abstracts of aerodynamic journal articles, a set of 225 queries and exhaustive relevance judgments from all pairs (query, document).
Table 1. Result of document indexing system

|          | Cran MAP | Cran MRR | NLP MAP | NLP MRR | CPSA MMAP | CPSA MRR |
|----------|----------|----------|---------|---------|-----------|----------|
| LSI-BOW  | 0.42     | 0.39     | 0.38    | 0.39    | 0.85      | 0.8      |
| LSI-TFIDF| 0.44     | 0.4      | 0.42    | 0.4     | 0.9       | 0.85     |
| LDA-BOW  | 0.37     | 0.35     | 0.35    | 0.14    | 0.66      | 0.67     |
| LDA-TFIDF| 0.43     | 0.36     | 0.40    | 0.37    | 0.7       | 0.75     |
| DOC2VEC  | 0.28     | 0.27     | 0.22    | 0.19    | 0.43      | 0.4      |

After analyzing the results of different approaches using MRR and MAP metrics, it was found that LSA representation based on TFIDF weighting is the most efficient, its performance is improved according to the increase of the size of the dataset.

3. A new automatic extractive summary approach

Automatic text summary consists in producing a concise and fluid summary while preserving the content of key information and its global meaning [4]. Among the methods of automatic summary of textual data, there are extractive abstract techniques that produce summaries by choosing a subset of sentences in the original text. These summaries contain the most important sentences of the entry. The entry can be a single document or several documents. [3]

3.1. Proposed method

We have proposed an Extractive Automatic Summary Approach that uses a subject modeling algorithm. The objective is to extract the sentences from a document that are most representative of its content.

To do this, we represent the document by the LSA model and then extract k dominant topics from the document where k varies according to the user's need, and then extract the most important sentences for each topic, Then for each average value sentence is calculated to choose among these most representative sentences of these topics that will constitute the summary.

![Figure 3. System Process Realized](image)

3.2. Experiment and Results

For the purposes of our study, the corpus used consists of 2225 documents from the BBC news site corresponding to articles in five thematic areas from 2004-2005. (Business, Entertainment, Politics, Sport, Technology). For each article, a summary is provided in the Abstracts folder.

we generated the summaries by the proposed method and compared it with the summaries generated from the MS Word Office package.
Table 2. Results for the proposed method

|          | BLEU  | ROUGE-1 | ROUGE-2 | ROUGE-3 | ROUGE-4 | ROUGE-L |
|----------|-------|---------|---------|---------|---------|---------|
| Accuracy | 0.432 | 0.771   | 0.53    | 0.491   | 0.468   | 0.668   |
| Rappel   |       | 0.684   | 0.47    | 0.433   | 0.41    | 0.621   |
| F*Measure|       | 0.713   | 0.497   | 0.456   | 0.43    | 0.643   |

Table 3. Results for Ms word summarizer

|          | BLEU  | ROUGE-1 | ROUGE-2 | ROUGE-3 | ROUGE-4 | ROUGE-L |
|----------|-------|---------|---------|---------|---------|---------|
| Accuracy | 0.430 | 0.566   | 0.492   | 0.487   | 0.451   | 0.607   |
| Rappel   |       | 0.645   | 0.419   | 0.412   | 0.391   | 0.608   |
| F*Measure|       | 0.721   | 0.389   | 0.412   | 0.4     | 0.512   |

In this experiment, we measured the similarity of summaries with human summaries in terms of the different variants of the ROUGE [15] and BLUE [5] scores. The correlation results are shown in Table 2 and 3.

We can observe that when comparing the summaries generated by our system with MS word summarizer, our approach proves good results. This experiment allowed us to affirm that the approach used in our system is competitive with the summaries generated manually by humans.

4. Improved indexing system

4.1. Proposed method

In order to minimize the execution time of the indexing phase and the memory used, we proposed to index the summaries instead of the original documents.

So, Summary Indexing allows you to quickly search large datasets by spreading the cost of a time-consuming document.

You can then quickly and efficiently search this small subset of data in the abstract index.

Figure 4. Summary indexing process
4.2. Experiment and Results

The results obtained by this system were evaluated for the CPSA data set.

![Indexing results by summary on CPSA](image)

**Figure 5.** Indexing results by summary on CPSA

We notice from Figure 5 that the accuracy has not changed but the recall is decreased, to properly verify the importance of the returned documents, the system was evaluated by MRR (Mean Reciprocal Rank) measurements which evaluate the average rank of the first relevant document in the results list and MAP (Mean Average Precision) and the following results were obtained:

**Table 4.** MAP and MRR results on CPSA

| Metric | Value |
|--------|-------|
| MAP    | 0.96  |
| MRR    | 0.98  |

It was noted that the values obtained by the improved indexing system for MRR and MAP metrics are higher than the values obtained by the normal indexing system, this implies that the hybrid system returns the most relevant documents at the top of the list of results.

**Table 5.** Comparison of execution times.

|                     | Run times       |
|---------------------|-----------------|
| Simple system       | 1.2499482631683350 s |
| Hybrid system       | 0.3437492847442627 s |

Table 4 shows the difference between the simple indexing execution time and the summary indexing time, we observe that the summary indexing system minimizes the execution time by 73% compared to the normal indexing system.

5. Conclusion

In this paper, we examined two areas of research in natural language processing, automatic indexing, and automatic summarization. A hybridization of a thematic indexing system and an automatic summarization system has made it possible to create an improved indexing system, which starts with a
summary phase of the document to be indexed before starting the indexing phase. The effectiveness of the combination performed is demonstrated by a set of experimental results obtained.

References
[1] Lancaster, Frederick W. 2003. Indexing and Abstracting in Theory and Practice. 3rd ed. London: Facet.
[2] Clough, P. and Stevenson, M. Developing A Corpus of Plagiarised Short Answers, Language Resources and Evaluation: Special Issue on Plagiarism and Authorship Analysis, In Press.
[3] Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In Mining Text Data. Springer, 43–76.
[4] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. In Proceedings of arXiv, USA, July 2017, 9 pages.
[5] Papinen, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei Jing. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation.
[6] Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. JASIS, 41(6):391–407, 1990.
[7] Blei DM, Ng AY, Jordan MI, Lafferty J. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003 05/15 ;3(4) :993-1022.
[8] Tonelli S., Cabrio, E. et Pianta, E. (2012). Key-concept extraction from french articles with kx. In Actes de l’atelier de clôture du huitième défi fouille de texte (DEFT), pages 19–28
[9] Pianta, E., Girardi, C. et Zanoli, R. (2008). The TextPro tool suite. In Proceedings of the 6th Language Resources and Evaluation Conference (LREC), Marrakech, Morocco.
[10] El Ghali, A., Hromada, D. et El Ghali, K. (2012). Enrichir et raisonner sur des espaces sémantiques pour l’attribution de mots-clés. In Actes de l’atelier de clôture du huitième défi fouille de texte (DEFT), pages 81–94.
[11] Cohen, T., Schvaneveldt, R. et Widdows, D. (2010). Reflective random indexing and indirect inference: A scalable method for the discovery of implicit connections. Biomed Inform, 43(2): 240–256.
[12] Morgane Marchand, Geoffroy Fouquier, Guillaume Pitel ,Représentation vectorielle de mots pour l’indexation de notices bibliographiques, Actes de la conférence conjointe JEP-TALN-RECIHAL 2016, volume 8 : DEFT
[13] Adrien Bouguin, Florian Boudin, Béatrice Daille, TopicRank en domaines de spécialité : participation du LINA à DEFT 2016, Actes de la conférence conjointe JEP-TALN-RECIHAL 2016, volume 8 : DEFT
[14] Bouguin A., Boudin F. & Daille B. (2013). Topicrank: Graph-Based Topic Ranking for Key phrase Extraction. In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP), p. 543–551, Nagoya, Japan: Asian Federation of Natural Language Processing.
[15] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using ngram co-occurrence statistics. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 71–78. Association for Computational Linguistics, 2003.