A SEMANTIC CLUSTERING APPROACH FOR INDEXING DOCUMENTS

Daniel Osuna-Ontiveros, Ivan Lopez-Arevalo and Victor Sosa-Sosa
Information Technology Laboratory, CINVESTAV - IPN, Av. Victoria-Soto La Marina Km. 5.5 Ciudad Victoria, Tamaulipas, Mexico

Keywords: Indexing models, Information retrieval, Semantic clustering, Semantic search.

Abstract: Information retrieval (IR) models process documents for preparing them for search by humans or computers. In the early models, the general idea was making a lexico-syntactic processing of documents, where the importance of the documents retrieved by a query is based on the frequency of its terms in the document. Another approach is return predefined documents based on the type of query the user make. Recently, some researchers have combined text mining techniques to enhance the document retrieval. This paper proposes a semantic clustering approach to improve traditional information retrieval models by representing topics associated to documents. This proposal combines text mining algorithms and natural language processing. The approach does not use a *priori* queries, instead clusters terms, where each cluster is a set of related words according to the content of documents. As result, a document-topic matrix representation is obtained denoting the importance of topics inside documents. For query processing, each query is represented as a set of clusters considering its terms. Thus, a similarity measure (e.g. cosine similarity) can be applied over this array and the matrix of documents to retrieve the most relevant documents.

1 INTRODUCTION

The actual increase of electronic information has become difficult the search of information by people. Some representation models as boolean representation or vector space model (Salton et al., 1975) (VSM) has been proposed to represent documents in order to make information readable by computers. The disadvantage of these models is their search is based only considering the terms of a query. Other models as probabilistic model (Robertson and Jones, 1976) and latent semantic indexing (Deerwester et al., 1990) (LSI) use a mathematical approach in order to find hidden relations between the terms in documents. Some modifications to these models have been suggested in order to improve their performance.

With the aim to enhance the above approaches, some text mining algorithms have been applied to get knowledge from documents.

Other approaches as Latent Dirichlet Allocation (Blei et al., 2003) (LDA) and Clusteing by Committee (Pantel, 2003) (CBC) were proposed in order to recover topics from a set of documents. These algorithms process documents to cluster terms where clusters can be seen as topics or a set of related terms. Griffiths and Steyvers (Griffiths and Steyvers, 2004) presented a statistical inference algorithm for LDA using scientist papers to test their proposal. In Bioinformatics, Konietzny *et al* (Konietzny et al., 2011) applied LDA to identify functional modules of protein families. The method explores the co-occurrence patterns of protein families across a collection of sequence samples to infer a probabilistic model of arbitrarily-sized functional modules.

Other approaches have been used to improve probabilistic information retrieval models. Lafferty and Zhai (Lafferty and Zhai, 2001) propose a framework for information retrieval that combines document models and query models using a function based on Bayesian decision theory. In their proposal, Lafferty and Zhai estimate a language model for each document, as well as a language model for each query, and the retrieval problem is cast in terms of risk minimization. Ponte and Croft (Ponte and Croft, 1998) presents a language modeling approach to improve the weighting proposed by Salton *et al* (Salton et al., 1975). They use the mean probability to terms as an estimator to model the relevance of a term in a document.

This paper proposes a semantic information re-

Osuna-Ontiveros D., Lopez-Arevalo I. and Sosa-Sosa V..
A SEMANTIC CLUSTERING APPROACH FOR INDEXING DOCUMENTS.
DOI: 10.5220/0003663802800285
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 280-285
ISBN: 978-989-8425-79-9

trieval model based on topics for document retrieval. The remain of the paper is structured as follows. Section 2 presents the background on information retrieval models. Section 3 describes the process to build the model representation. Section 4 shows the results of an experiment to demostrate the performance of the model. Finally, some conclusions and remarks are given in Section 5.

2 BACKGROUND

This section shows the theoretical basis of this work and some proposals reported in the literature.

2.1 Information Retrieval

The recovery and representation of information is defined as the model, design, and implementation of systems to provide quick and effective access to the contents of documents (Manning et al., 2008). The purpose of information retrieval systems is to represent documents in order to estimate its relevance based on the user's search.

Some of the most used approaches for document representation are:

- *Boolean Representation* (BR) is a classical model for information retrieval which documents and terms are represented by an incidence matrix. The order of terms is not relevant in this model. Only important is to know whether a term is found or not in a document. Using the boolean representation is easy to know whether the terms of a query are in a document. The disadvantage of this model is its inability to know which document is more relevant with respect to the query.
- The *Vector Space Model* (VSM) (Salton et al., 1975) is based on the boolean representation. The vector space model uses a *document-term* matrix to represent the importance of each term in each document. This importance is based on the frequency of terms in documents. The $TF \cdot IDF$ is the weight often used to normalize this matrix. TF (Term Frequency) and IDF (Inverse Document Frecuency) are shown in Equations 1 and 2, respectively:

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2}$$

where $n_{i,j}$ is the number of occurrences of the considered term t_i in document d_j, $\frac{n_{i,j}}{\sum_k n_{k,j}}$ is the sum of number of occurrences of all terms in document d_j and $|\{j : t_i \in d_j\}|$ represents the number of documents where the term t_i appears.

- The *Probabilistic Model* (Robertson and Jones, 1976) uses statistical techniques to assign weights to documents. This model obtains relevant and irrelevant documents, and based on the obtained documents sorts them taking into account their importance.
- *Latent Semantic Indexing* (LSI) (Deerwester et al., 1990) is an information retrieval method which attempts to capture this hidden structure by using techniques from linear algebra. Vectors representing the documents are projected in a new low-dimensional space obtained by singular value decomposition of a *term-document* matrix A. This low-dimensional space is spanned by the eigenvectors of $A^T A$ that corresponds to the few largest eigenvalues and to the hidden correlations between terms. Queries are also projected and processed in this low-dimensional space.

2.2 Text Mining

Text mining is a branch that emerges from data mining. This branch seeks for knowledge into text documents. Text mining is used in information retrieval, document summarization, categorization of documents, clustering of terms/documents, etc. The following text mining algorithms are some of the most relevant in this work.

- *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003) is an algorithm to obtain the most representative terms within a *corpus*. LDA is a probabilistic approach based on a bayesian model for associating words to topics. This method is based on the idea that each document refers to a number of topics and using probability models defines the belongings of each term respect to topics. LDA can be used for terms reduction and disambiguation. Terms that are not associated to any topics are considered irrelevant to the *corpus*. LDA get topics as follows:

  1. Choose $\theta_i \sim Dir(\alpha)$, where $i \in \{1,\ldots,M\}$
  2. For each of the words w_{ij}, where $i \in \{1,\ldots,N_i\}$
  (a) Choose a topic $z_{ij} \sim Multinomial(\theta_i)$
  (b) Choose a topic $w_{ij} \sim Multinomial(\beta_{z_{ij}})$

  where:

  – α is the parameter of the uniform Dirichlet prior on the per-document topic distributions.
  – β is the parameter of the uniform Dirichlet prior on the per-topic word distribution.

- $theta_i$ is the topic distribution for document i,
- z_{ij} is the topic for the jth word in document i, and
- w_{ij} is the specific word.

- *Clustering by Committee* (CBC) (Pantel, 2003) is an algorithm used to cluster terms. CBC organizes documents by topics to discover concepts and meaning of words. The difference with LDA is that CBC analyzes relations of terms (e.g., verb-noun), while LDA only process terms. It has been used to get related terms based on a set of documents (Lin, 1998). CBC work as follows:

1. For each element are obtained the most similar elements to create n clusters.
2. It is obtained a committee list
   (a) For each element, cluster the top similar elements from similarity database using average-link clustering.
   (b) For each cluster: add centroid's to committee if centroid's similarity to the centroid of each committee previously added is below a threshold θ_1
   (c) Add element to a list of residues if similarity is below by threshold θ_2.
   (d) When list of residues is empty, return committee.
3. Groups are created, where the committees created in the previous phase are the centroids of these groups.

3 METHODOLOGY

This section presents the steps to build the proposed *document-topic* matrix. The first part of the methodology raises the representation of information. The second part is to process the search of a user. The steps are illustrated in Figure 1.

3.1 Document Representation

First, it is necessary to obtain the contents from documents to be processed. A good practice is to discard *stopwords* which are words that do not contribute to documents. The proposed representation model only considers verb-noun relations for processing. Nouns that do not have associated a verb provide little semantic importance related to the document. For this task, in the implementation, the Stanford tagger (Klein and Manning, 2003) is used. Verb-noun relations are obtained for each input file for integrating all relationships in one file. The file with all relationships is the input to a modified version of the CBC algorithm. With *CBC*, the terms can be grouped based on verb-nouns relationship. The proposed approach assumes that groups generated by CBC are topics and the similarity of a term respect to its centroid corresponds to the relevance of a term to the topic. This will create a topic-term matrix that represents the importance of terms for each topic. The matrix is represented as shows Table 1.

Table 1: *Topic-term* matrix.

Topic \ Term	Term-1	Term-2	Term-3	Term-4	Term-5
Topic-1	0.102	0.016	0.000	0.123	0.000
Topic-2	0.000	0.012	0.130	0.100	0.030
Topic-3	0.063	0.002	0.023	0.102	0.002
Topic-4	0.030	0.123	0.020	0.021	0.015

Subsequently, LDA is applied to cluster terms. Terms that were not added to any cluster (outliers) and *stopwords* are removed from documents. In this approach, every document is seen as a *bag of words*. With this set of *bag of words*, it is created a frequency matrix. A frequency matrix denotes the frequency that terms appears in documents as is shown in Table 2.

Table 2: *Term-document* matrix.

Term \ Document	Doc-1	Doc-2	Doc-3
Term-1	1	4	0
Term-2	2	2	4
Term-3	3	3	2
Term-4	0	0	2
Term-5	4	2	3

Then, a *topic-document* matrix (γ) is obtained as the product of the *topic-term* matrix (α) and the *term-document* matrix (β) as shows Equation 3

$$\gamma = \alpha \cdot \beta \tag{3}$$

The *topic-document* matrix is transposed to get a *document-topic* matrix as shows Table 3. The idea to transpose the matrix is for facilitating the processing of queries. This matrix is stored as an index.

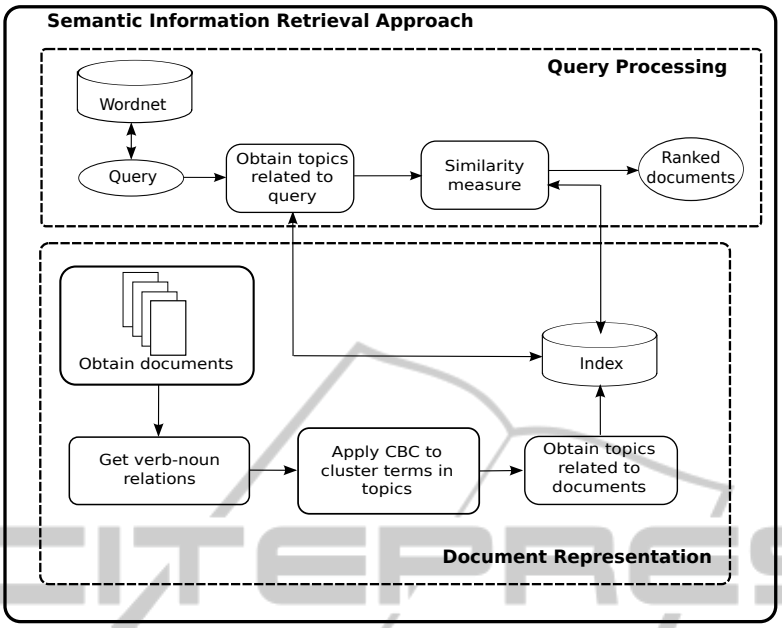

Figure 1: Proposed information retrieval approach.

Table 3: *Document-topic* matrix.

Document \ Topic	Topic-1	Topic-2	Topic-3	Topic-4
Doc-1	0.343	0.232	0.238	0.264
Doc-2	0.440	0.474	0.329	0.456
Doc-3	0.340	0.598	0.264	0.619

3.2 Query Processing

In a similar way as the *document-topic* matrix was created, the terms of a query are processed to obtain its relevance to each topic. The goal is to create a vector that represents the topics of the query. For example, the "relation" of the query terms *term-1* and *term-2* (both terms) to each document can be seen as shows Table 4.

Table 4: *Document-topic* matrix with the query represented as a topic vector.

Document \ Topic	Topic-1	Topic-2	Topic-3	Topic-4
Doc-1	0.343	0.232	0.238	0.264
Doc-2	0.440	0.474	0.329	0.456
Doc-3	0.340	0.598	0.264	0.619
Query	0.160	0.142	0.025	0.143

In this way, the cosine similarity measure can be applied over the vectors from the *query-topic vector* (A) and the set of *document-topic vectors* (B) for retrieving the most representative documents.

4 PRELIMINARY RESULTS

We conducted an experiment to evaluate the approach making 31 queries (according to the Central Limit Theorem (Fischer, 2011)) over the *reuters* corpus. This corpus has 12902 documents about 116 topics. The results were compared against BR and VSM models. It was used the precision, recall, and f-measure. For the precision, were recovered the 25 most relevant documents (according to the study of Sanchez-Ruenes (Sánchez, 2009)) for each query in order to measure the ranking of precision in models. For the recall, a treshold of 0.2 was used in VSM and the proposal in order to measure the ranking of recall in models.

It is important to remark that queries are based on the idea of topic-based search. A resulting document is considered relevant although it do not contains a specific query-term. For example, a query like "steel" could returns documents with information related to steel, as "iron" or "metal", which are very related terms to "steel".

In Figures 2, 3 and 4, can be seen from the results, that in general, this approach performs with higher effectiveness than BR and VSM. Table 5 shows the average of precision, recall and f-measure obtained by models.

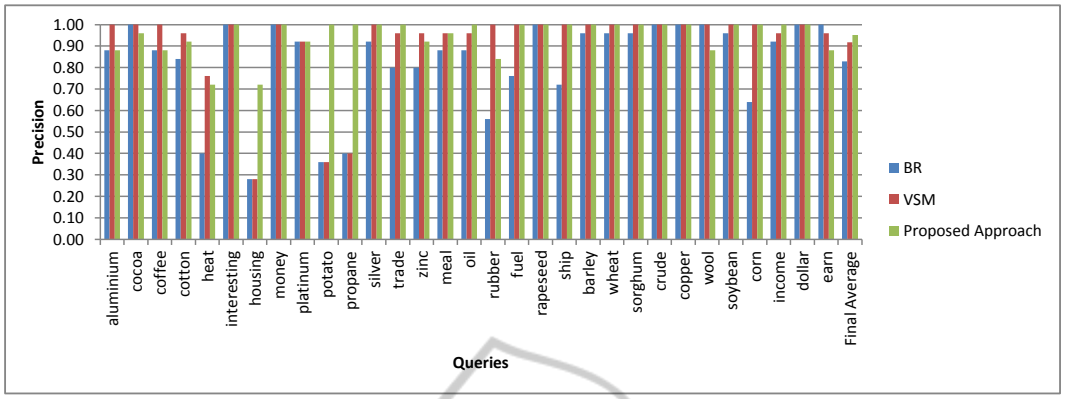

Figure 2: Precision of 31 queries using BR, VSM, and the proposal.

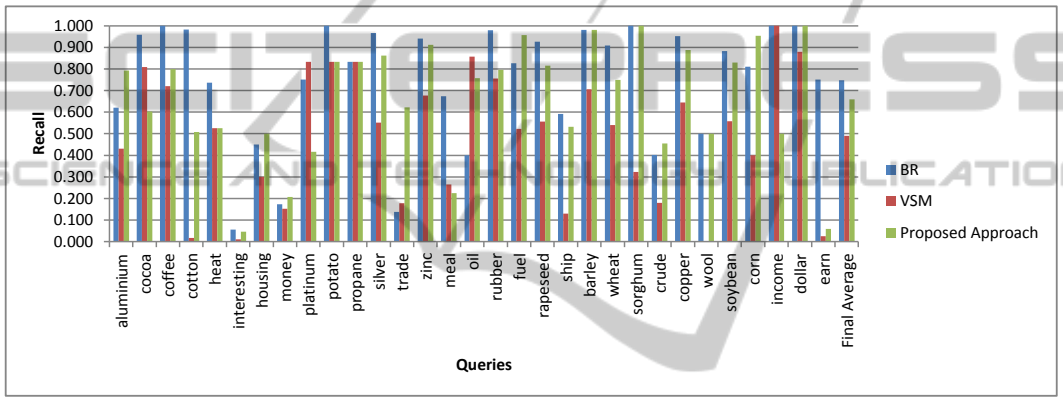

Figure 3: Recall of 31 queries using BR, VSM, and the proposal.

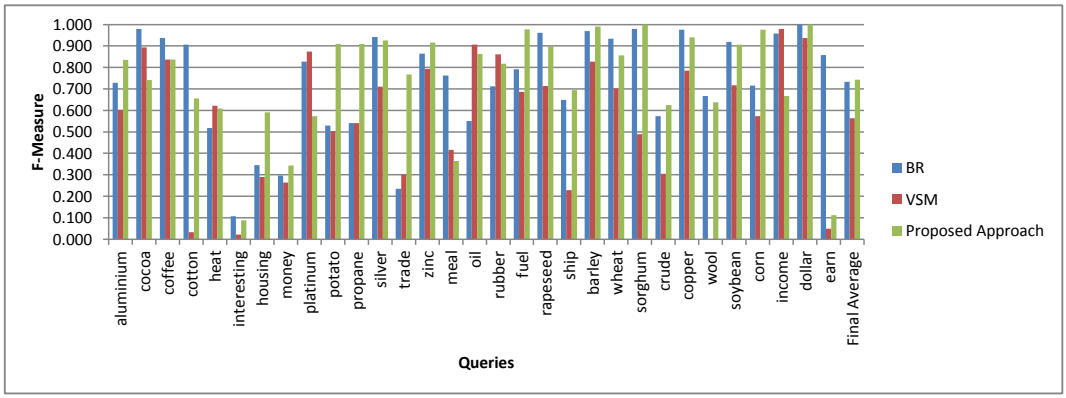

Figure 4: F-Measure of 31 queries using BR, VSM, and the proposal.

Table 5: Average of precision, recall, and, f-measure over BR, VSM, and the proposal.

Metric	BR	VSM	Proposed Approach
Precision	0.828	0.917	0.951
Recall	0.748	0.491	0.660
F-Measure	0.733	0.563	0.743

Moreover, for specific queries, the proposed model had a regular behavior. From the precision, was seen that while BR and VSM returned few documents (10 or less) for a query, the proposed approach returned more results with relevance (more than 20). This is because in the proposed IR model is not nec-

essary that a document contains terms of query. From the recall, the boolean representation has the better result. This was because BR returned a lot of documents (more than 300) for each query.

5 CONCLUSIONS

An unsupervised approach for indexing documents is proposed in this paper. The proposal combines text minig and natural language processing to obtain a document-topic matrix representation for a set of documents. First, verb-noun relationships are obtained by using a POS tagger. Then the Clustering by Committee algorithm is used to group terms according to verb-noun relations. After that, the Latent Direchlet Allocation is applied to obtain the most relevant terms. The parameters for LDA are obtained without human intervention. According to the experiments, in general, the proposal has a better performance in topic-based semantic searches over traditional models (boolean and vector space model). Future work should include semantic processing in web search or analysis of tweets/posts.

ACKNOWLEDGEMENTS

This research was partialy funded by project number 165474 from "Fondo Mixto Conacyt-Gobierno del Estado de Tamaulipas".

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41:391–407.

Fischer, H. (2011). Conclusion: The central limit theorem as a link between classical and modern probability theory. In *A History of the Central Limit Theorem*, Sources and Studies in the History of Mathematics and Physical Sciences, pages 353–362. Springer New York.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science*, 101:5228–5235.

Klein, D. and Manning, C. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*.

Konietzny, S. G. A., Dietz, L., and McHardy, A. C. (2011). Inferring functional modules of protein families with probabilistic topic models. *BMC Bioinformatics*, 12:141.

Lafferty, J. D. and Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In Croft, W. B., Harper, D. J., Kraft, D. H., and Zobel, J., editors, *SIGIR*, pages 111–119. ACM.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA. Association for Computational Linguistics.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Pantel, P. A. (2003). *Clustering by committee*. PhD thesis, University of Alberta Edmonton. Adviser-Dekang Lin.

Ponte, J. and Croft, B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st International Conference on Research and Development in Information Retrieval*.

Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620. The paper where vector space model for IR was introduced.

Sánchez, D. (2009). Domain ontology learning from the web. *The Knowledge Engineering Review*, 24(04):413–413.