International Arab Conference on Information Technology (ACIT'2016)

A Hybrid Approach of Semantic Similarity Calculation for a Content-based Recommendation of Text Documentson an E-learning Platform

Badr HSSINA[1], Abdelkrim MERBOUHA[2], and Belaid BOUIKHALENE[1]
[1]TIAD Laboratory, Computer Sciences Department Sultan MoulaySlimane University, FST
Beni-Mellal, Morocco
[2]LMACS Laboratory, Mathematics Department Sultan MoulaySlimane University, FST
Beni-Mellal, Morocco

Abstract:*Currently in theonline learning sector, electronic monitoring of items is crucial. Maintaining effective monitoring involves targeting items to consult with users because the information amount is important. To resolve this problem, we propose an innovative recommendation system of documents which is based on the integration of its semantic indexing. In this context, we have created a semantic similarity calculation system between text documents to help their semantic recommendations. Indeed, the semantic recommendation of documents is a promising field of research, because it guarantees a quick and targeted access to information. The aim of our work is to guide learners and suggest resources on the basis of their learning experience. Our approach is to build a semantic recommendation system based on content; it is a system that allows returning from a set of documents which isrelevant to a learner--that is to say, the documents that are semantically similar to a document chosen by the learner. Experimental evaluations using WordNet prove that our system improves the accuracy of the semantic recommendation text documents to learners.*

Keywords:*Recommendation system, semantic similarity, WordNet.*

1. Introduction

Nowadays, recommendation systems have become an independent research field[1]. Indeed, with the development of Web and especially e-Learning platforms, interest in recommender systems has increased significantly. The original recommendations of the systems appeared to try to resolve the problems associated with information overload (cognitive overload)[2].

Recommender systems are tools that can provide personal recommendations or to guide users to interesting or useful resources within a large data space [3]. Recommender systems are primarily geared towards individuals who do not have enough personal experience or expertise to assess the potentially huge amount of information that a Web site can offer as an example [4].

The two basic elements that appear in every recommendation systems are theitem and the user:

- Item is the general term used to denote that which the system recommends to users. To offer useful and relevant suggestions for a particular user, a recommendation system will focus on a specific type of items (for example: web pages, documents, pictures, etc.) and therefore will customize its navigation model, its interface and its technical recommendation.
- Users should be modeled in the recommendation system so that the system can exploit their profiles and preferences. Moreover, a clear and precise description of the contents is also required to obtain good results at the time of recommendations [5].

It is possible to classify therecommendation systems in different ways. In the most known categorization, there are two approaches: the recommendation based on the content and the recommendation by collaborative filtering [6] [7].

The recommendation based on content [8]: The system recommends items that are similar to those that the user has liked in the past. The similarity of the items is calculated based on the characteristics associated with the compared items. For example, if the user has positively noted a show that belongs to the comedy genre, the system can provide this kind of entertainment recommendations.

The recommendation by collaborative filtering [9]: The system asks users to evaluate resources, so they know what they like most. Then, when a recommendation is requested for the current user, the system will suggest resources that similar users have already liked. Collaborative filtering is the most popular and most widespread technique in recommender systems.

A recommendation system can be quite complicated to develop. In addition to be effective, there must be a certain number of users, have collected a good amount of information and use appropriate algorithms.

2.Related Works

In this part of the state of the art, we emphasize on the recommendation based on content, calculating the semantic similarity between text documents, by calculating similarity between the items. Systems based on semantics are a special case of content-based systems. They tend to integrate new technologies of the Semantic Web [10] in order to remedy certain shortcomings of the conventional content-based systems.

The first recommendation system has adopted a performance based on the meaning of documents to build a model of the user's interests was SiteIF [11]. This is a personal agent for a website multi-lingual. The external knowledge source involved in the process of representation is MultiWordNet [16], a multi-lingual lexical database. Each concept is automatically associated with a list of sets of synonyms (called synsets). The user profile is a semantic network in which nodes represent synsets found in the documents read by the user. During the correspondence phase, the system receives as input the representation of synsets of a document and the current user model, and outputs an estimate of the relevance of the document using the Semantic Network Value Technique [12].

ITR (ITerm Recommender) [13] is a system capable of providing recommendations in several areas of items (movies, music, books), provided that the article descriptions are available in text documents. ITR incorporates linguistic knowledge in the learning process of users' profile. The linguistic knowledge comes exclusively from the WordNet lexical ontology [17]. The items are represented as vectors based on synsets. A vector of synsets, rather than a vector of words corresponds to a document. The user's profile is constructed as a binary text classifier naive Bayesian [14], to categorize an item as interesting or not. It includes the synsets that prove to be more indicative of user preferences, depending on the value of the conditional probabilities estimated in the learning phase. The item-profile match is to calculate the probability that an item belongs to the class interesting, using probabilities of synsets in the user's profile.

JUMP [15] is a system able to deliver intelligently contextual information, personalized to users. User's needs are represented as complex queries, rather than as a user profile. Semantic analysis of documents and analysis of user needs is based on domain ontology where each concept is manually annotated using WordNetsynsets. The correspondence between the documents and the field of concepts is done automatically by the procedures that exploit lexical annotations in the domain ontology. The main interest for the language skills is emphasized by the wide use of WordNet, which is mainly used for the interpretation of the content using disambiguation. On the other hand, the previously described studies show that WordNet is not sufficient for full understanding of user interests and their contextualization. Specific knowledge of the domain and the context is necessary.

3. Content-basedRecommendationSystem

Recommendation systems based on the content (Figure 1) are at the intersection of the fields of information retrieval systems and artificial intelligence [18]. The recommendation technique based on the contents is based on the assumption that the items with similar content will be equally appreciated [19]. This technique is based on analyzing the content of similarities between the items previously viewed by users and those who have not yet been consulted [3]. Items that can be recommended to users are represented by a set of characteristics, also called attributes, variables or properties in the literature. For example, in an application for recommending films, attributes adopted to describe a film are: actors, directors, genres, subject, etc., or in an e-learning platform adopted attributes to describe the content are: title, author, subject, etc. An item is represented in the system by means of a structured data. More formally, this structured data is a vector$X = (x1, x2, \cdots, xn)$of n components, where each component represents an attribute.

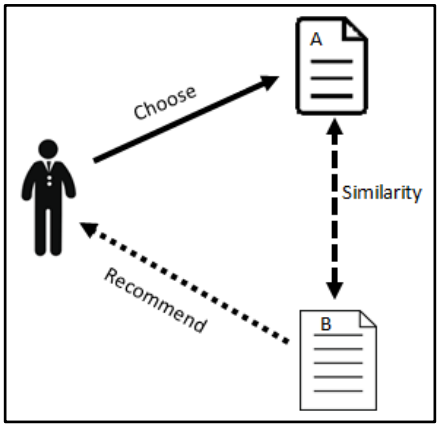

Figure 1. Recommendation based on content

Most content-based recommendation systems use simple research models originally used for research information, such as correspondence of keywords or the Vector Space Model (VSM) [20], and couples with basic weighted TF-IDF [21]. VSM is a spatial representation of textual records and the TF-IDF measure is a statistical measure that evaluates the significance of a word in a document or an item that is part of a collection of documents [8]. In this model, each document is represented by a n-dimensional vector, where each dimension corresponds to a term of the entire vocabulary of a collection of documents. Formally, any document is represented by a vector weight on words, or each weight indicates the degree of association between the document and the word: Let$D = \{d1, d2, \cdots, dN\}$denoting a set of N records of corpus, and$T = \{t1, t2, \ldots, tn\}$ is the dictionary or all words in the corpus. T is obtained by applying automatic processing operations of natural language (TAL) [22], tokenization [23], elimination of stop words, stemming and [18]. Figure 2 illustrates our proposed approach to extract the terms of the corpus.

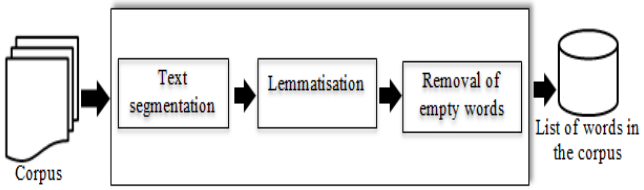

Figure 2. Proposed Approach to extract the terms of the corpus

Each document dj is represented by a vector in a vector space of n dimensions,
as$j = \{w1j, w2j, \ldots, wnj\}$, where$w_{kj}$ is the weight of the termt_k in the documentd_j obtained with calculating TF-IDF (term Frequency-InverseDocument Frequency) [24], the $tfidf(t, A, D)$ weightof the term t, belonging to the corpus D of document A is obtained:

$$tfidf(t, A, D) = \frac{n_{t,A}}{N_A} \times \log\left(\frac{|D|}{|\{A \in D : t \in A\}|}\right)$$

Where:

- $n_{t,A}$is the number of occurrence of t inA andN_d the document size of A.
- $|D|$: total number of documents in the corpus;
- $|\{A \in D : t \in A\}|$: number of documents where the term t appears.

This is to divide the total number of documents in the corpus D by the number of documents containing the term t, and calculate the logarithm. The inverse document frequency (Inverse Document Frequency) is a measure of the importance of the word in the whole corpus. The TF-IDF aims to give greater weight to the least frequent words, considered more discriminating.

4. Architecture of our system

In this work, we propose a system for calculating the similarity between text documents. The first phase is indexing and semantic enrichment from the WordNet lexical database and after TF × IDF weighting for representing documents in the vector model. All these operations must be validated before we apply our similarity computing approach. Finally, we recommend to learners similar documents to the reference document based on the similarity values obtained (Figure 3).

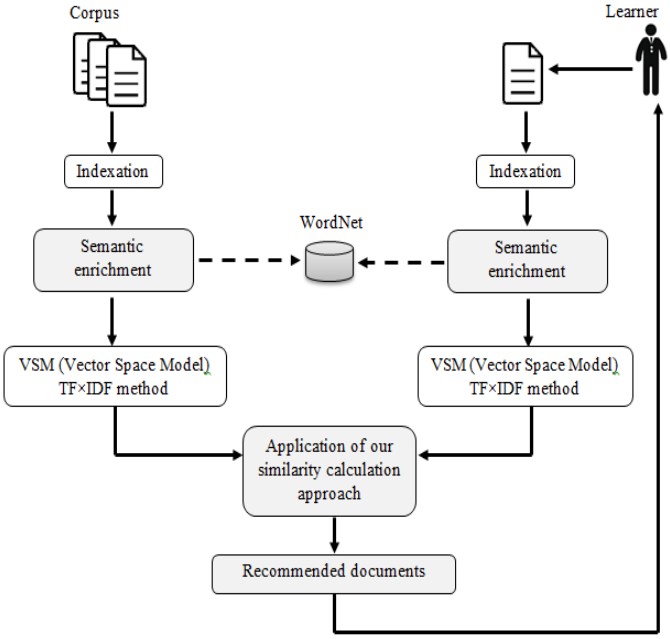

Figure 3.Our approach to calculation of similarity between the text documents for a semantic recommendation

We apply our approach to build a semantic recommendation system based on content. It is a system which allows searching for the documents

which are relevant to a user that are semantically similar to a document reference chosen by the learner.
The semantic enrichment is a key step in this work. We introduce a semantic relationship which is synonymy based on an external language resource WORDNET. The idea is to enrich the vector representation by the synonyms.
We are looking for synonymsfor each word of the vector representation of the document, after we investigate whether these synonyms are in the document, then if it is found we add its frequency in the vector representing the document.

.

5. Comparison of four similarity measures used in textual analysis

A measure of semantic similarity is a concept whereby a set of documents or terms are given a metric based on the similarity of their meaning / semantic content.

Specifically, this can be achieved by defining a topological similarity, for example, using ontologies for defining a distance between the words, or by defining a statistical similarity, for example the vector space model to correlate the terms and contexts from an appropriate text corpus.

Word similarity approaches based on knowledge is based on a semantic network of words, such as WordNet. Given two words, their similarity can be estimated from their relative position in the hierarchy of the knowledge base. These similarity measures include: Wu and Palmer [25], Resnik [26] Jiang Conrath [27] Lin [28].

To select an approach, we try to calculate the semantic similarity between identical documents, then we choose the approach that gives good results. Table 1 shows the results.

Table 1.Comparison of different methods of calculation of similarity in terms of execution time.

	similarité	Temps en milliseconde
Resnik	0.05	1461
Wu and Palmer	0.14	1399
Lin	0.03	1445
JianConrath	0.05	1492

From this table, it follows that the best approach is the approach of Wu and Palmer which gives the greatest similarity with a minimum running time. Based on the results of this comparison, we choose to work in our system with the Wu and Palmer method to calculate the similarity between two concepts i and j.

The returned document is sorted by a decreasing order of semantic similarity after applying a recommendation threshold to this similarity.If the value of the similarity is greater than or equal to the threshold, the recommendation system will return it. If not, the system will reject it..

6.Calculation of similarity

Since the objective is to evaluate the semantic similarity between documents, because these documents are represented by vectors (concept, weight), we use a technique to measure the similarity between weighted sets of concepts. After indexing each document, one can calculate the semantic similarity between two documents on the basis of these approaches, but this time not only between words but between documents which contains a set of words. Our approachcalculates semantic similarity between two documents (d and q) with $\vec{d} = (d1, d2, \dots dn)$and $\vec{q} = (q1, q2, \dots qn)$is defined by the following formula:

$$Sim(d,q) = \frac{\sum_i \sum_j d_i\, q_i \times sim(i,j)}{\sum_i \sum_j d_i q_j}$$

Where:
i : represents the concepts of the document q.
j : represents the concepts of the document d.
q_i: is the weight of i the concept in the document q.
d_j: is the weight of j the concept in the document d.
$sim(i,j)$ is the semantic similarity between the two concepts i and j, computed from the measurement of*Wu and Palmer*.

7. Experience and Evaluation

The principle of our application is to find from a collection of textual records (Corpus) documents that are semantically similar (relevant) for a document selected by the user of the application.

The quality of a system is to be measured by comparing the responses of the system with ideal responses that the user expects to receive. The more the system responses are consistent with those that the user hopes, the more the system is performing.
These two measures are given by the following formulas:
The recall is the ratio between the number of selected relevant items and the total number of relevant items.

$$\text{recall} = {}^{N_{ps}}\!/_{N_p}$$

Precision is the ratio between the number of selected relevant items and the total number of selected items.

$$\text{Precision} = {}^{N_{ps}}\!/_{N_s}$$

For example, the calculation of precision and recall of our recommendation system for Doc_1 and Doc_2 document is illustrated in Figure 4.

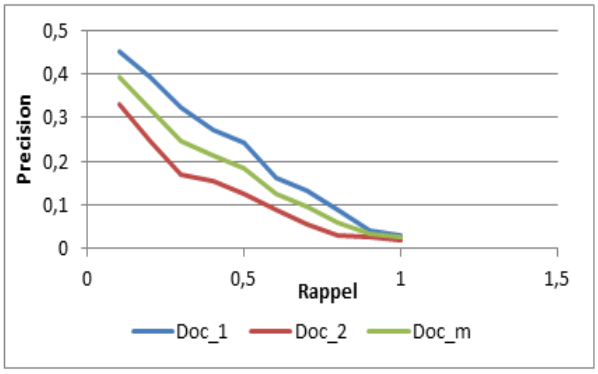

Figure 4.The recall-precision curve

To evaluate our semantic recommendation system from the system based on Cosine similarity, one must trace the Recall / Precision Curve for both SRI in the same graph (Figure 5).

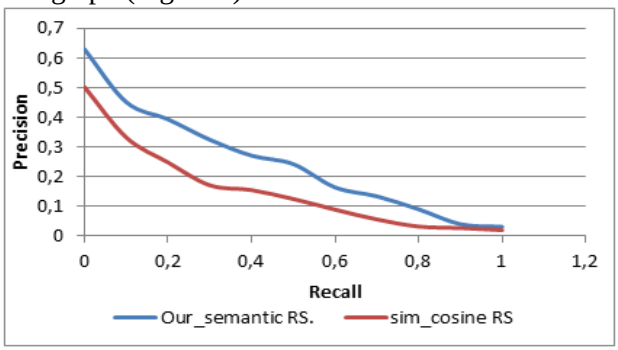

Figure 5.Recall/precisionCurve for our recommendation system and the one based on the cosine similarity

Before commenting on this result and as a reminder, we can say that if we want to compare two recommendation systems we must tested them with the same test corpus (or more test corpus). A system whose curve exceeds that of another is regarded as a better system. So according to this definition and from the curve of Figure 5, we conclude that our recommendation system is better than the system based on Cosine.

8. Conclusion

The current trend of recommendation systems is more focused on new methods, multi-criteria, multidimensional or based on psychological concepts such as emotions and opinions. Content-based recommendation systemsbuild document classes and offer to users the items from these classes, depending on the proximity of its profile with respect to classes. In this work, we have presented the steps to build our semantic recommendation system and how to achieve its evaluation using Precision and Recallmeasures.We have focused onmethods of semantic similarity calculation which is based onthe WordNet, lexical database, and the VSM. These methodsconsider the descriptors and weighting extracted from ourtextual corpus.

References

[1] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12), 61-70.

[2] Ménard, M. (2014). Systèmes de recommandation de biens culturels. *Les Cahiers du numérique*, *10*(1), 69-94.

[3] Burke, R. (2002). Hybrid recommender systems: Survey and experiments.User modeling and user-adapted interaction, 12(4), 331-370.

[4] Resnick, P., & Varian, H. R. (1997). Recommender systems. Communications of the ACM, 40(3), 56-58.

[5] Vozalis, E., &Margaritis, K. G. (2003, September). Analysis of recommender systemsalgorithms. In The 6th Hellenic European Conference on Computer Mathematics & its Applications (pp. 732-745).

[6] Shahabi, C., Banaei-Kashani, F., Chen, Y. S., & McLeod, D. (2001, September). Yoda: An accurate and scalable web-based recommendation system. In International Conference on Cooperative Information Systems (pp. 418-432). Springer Berlin Heidelberg.

[7] Picot-Clémente, R. (2011). Une architecture générique de Systèmes de recommandation de combinaison d'items: application au domaine du tourisme(Doctoral dissertation, Université de Bourgogne).

[8] Pazzani, M. J., &Billsus, D. (2007). Content-based recommendation systems. In The adaptive web (pp. 325-341). Springer Berlin Heidelberg.

[9] Lumineau, N. (2003). Un tour d'horizon du filtrage collaboratif. Rapport technique, LIP6, Paris, France.

[10] Berners-Lee, T., Hendler, J., &Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.

[11] Magnini, B., &Strapparava, C. (2004). User modelling for news web sites with word sense

based techniques. User Modeling and User-Adapted Interaction,14(2-3), 239-257.

[12] Stefani, A., &Strappavara, C. (1998, June). Personalizing access to web sites: The SiteIF project. In Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia HYPERTEXT (Vol. 98, pp. 20-24).

[13] Degemmis, M., Lops, P., &Semeraro, G. (2007). A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. User Modeling and User-Adapted Interaction, 17(3), 217-255.

[14] Poirier, D., Fessant, F., & Tellier, I. (2010). De la Classification d'Opinion à la Recommandation: l'Apport des Textes Communautaires. TraitementAutomatique des Langues, 51(3), 19-46.

[15] Basile, P., Degemmis, M., Gentile, A. L., Iaquinta, L., &Lops, P. (2007). The JUMP project: Domain Ontologies and Linguistic Knowledge@ Work. InSWAP.

[16] Gliozzo, A. M., Ranieri, M., &Strapparava, C. (2005, February). Crossing parallel corpora and multilingual lexical databases for WSD. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 242-245). Springer Berlin Heidelberg.

[17] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

[18] Baeza-Yates, R., &Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.

[19] Schafer, J. B., Frankowski, D., Herlocker, J., &Sen, S. (2007). Collaborative filtering recommender systems. In The adaptive web (pp. 291-324). Springer Berlin Heidelberg.

[20] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[21] Aizawa, A. (2003). An information-theoretic perspective of tf–idfmeasures.Information Processing & Management, 39(1), 45-65.

[22] Tanguy, L. (1997). Traitement automatique de la langue naturelle et interprétation: contribution à l'élaboration d'un modèle informatique de la sémantique interprétative (Doctoral dissertation, Université de Rennes 1).

[23] Mcnamee, P., & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. Information retrieval, 7(1-2), 73-97.

[24] Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning.

[25] Wu, Z., & Palmer, M. (1994, June). Verbs semantics and lexical selection. InProceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics.

[26] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007.

[27] Jiang, J. J., &Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

[28] Lin, D. (1998, July). An information-theoretic definition of similarity. In ICML(Vol. 98, pp. 296-304).