Assessing Novelty of Research Articles Using Fuzzy Cognitive Maps

S. Sendhilkumar, G.S. Mahalakshmi, S. Harish, R. Karthik,
M. Jagadish, and S. Dilip Sam

Abstract. In this paper, we compare and analyze the novelty of a scientific paper (text document) of a specific domain. Our experiments utilize the standard Latent Dirichlet Allocation (LDA) topic modeling algorithm to filter the redundant documents and the Ontology of a specific domain which serves as the knowledge base for that domain, to generate cognitive maps for the documents. We report results based on the distance measure such as the Euclidean distance measure that analyses the divergence of the concepts between the documents.

1 Introduction

We are facing an ever increasing volume of research publications. These have brought challenges for the analysis of novelty in these texts. Our main objective is to rate and find the novel information present in a journal of a specific domain. A large set of documents (corpus) of a specific domain is maintained so that the input journal of that domain is compared with those journals/documents to get the novelty score. The documents that are similar to the input document are found. After the similar documents are found, all those documents along with the input document are subjected to mapping by referring over Knowledge Base (Ontology) of that domain. The generated maps are Fuzzy Cognitive Maps (FCM) and they contain the required information to perform novelty computation. The corresponding maps of two documents are compared to find the divergence between two documents. To find the novel regions/parts of the input document, we have proposed a new measure to calculate the novelty score. Finding novelty in the scientific documents requires a knowledge base of a specific domain to analyse the concepts present in the documents. We used Ontology as the knowledge base for the domain to know the hierarchy of concepts and their relationships.

S. Sendhilkumar
Department of Information Science & Technology, Anna University, Chennai

G.S. Mahalakshmi · S. Harish · R. Karthik · M. Jagadish · S. Dilip Sam
Department of Computer Science & Engineering, Anna University, Chennai

A. Abraham and S.M. Thampi (Eds.): Intelligent Informatics, AISC 182, pp. 73–79.
springerlink.com

2 Related Work

2.1 Sentence-Level Approach

Sentence Level Novelty Detection aims at finding relevant and novel sentences given a query/topic and a set of documents. The cosine similarity is the widely used metric in the sentence level approach [9]. In Xiaoyan Li and Bruce Crofts work on Answer updating approach to Novelty Detection [2], they treated new information as new answers to questions that represented user's information requests (query). In Flora S. Tsais work on Novelty Detection for text documents [5], the named entities are assigned weights by using two different metrics, If the number of unique entities exceeded a particular threshold, the sentence was declared as novel. The other model such as vector space model [6] [1], Graph based text representation [4] ,Language model [3] ,Overlap relations [13] etc. are implemented for sentence level Novelty detection.

2.2 Document-Level Approach

In Zhang et als work on novelty and redundancy detection in adaptive filtering[12], the cosine metric and a mixture of probabilistic language models which is shown to be effective are used. Flora S. Tsais proposed D2S: Document-to-sentence framework for novelty detection [8] in which the novelty score of each sentence is determined to compute the novelty score of the document based on a fixed threshold.

3 System Design and Implementation

3.1 Distributional Similarity

By using LDA, the topic distributions over a number of documents are obtained. In this model each document d is represented by a topic distribution Θ_d. Kullback-Leibler divergence is a non-symmetric or distributional similarity measure of the difference between two probability distributions P and Q. Here it is used to the difference between topic distributions of two documents [12] which is one way to measure the redundancy of one document given another.

$$R(dt|di) = KLDiv(\Theta_{dt}, \Theta_{di}) = \sum(t_i|\theta_{dt})\log(P(t_i|\theta_{di})/P(t_i|\theta_{dt}) \quad (1)$$

Where R is the redundancy of document dt over document di and Θd is the topic model for document d and is a multinomial distribution. The documents that have least divergence value with the input document are selected so that redundant documents are filtered out and the documents that are more similar to the input document are selected for further processing.

3.2 Ontology Mapping

We view the ontology as the concept tree of that domain where each node represents a concept and their parent and child nodes represent their generalized and specialized form respectively. The concepts information include the level information of the concept in the tree i.e. height and depth of the concept in the tree, the occurring concepts (term) frequency in the document, the path to the root node of the concept etc. based on the needs of the novelty score computation. The map-ping process is per formed for the input document and for each document that is found during the similarity process and corresponding mappings are generated in xml format.

3.3 Divergence Analysis

The hopping distance between two nodes (concepts) represents the closeness of concepts to each other. For example, the hopping distance between the "RSSI measures" and "Route Request process" is 2 which have a common parent node "Route discovery". For further computation, Concept matrices are constructed for each document where the rows and columns of each matrix are the concepts found in both the documents and their corresponding matrix values are the hopping distance values between the nodes.

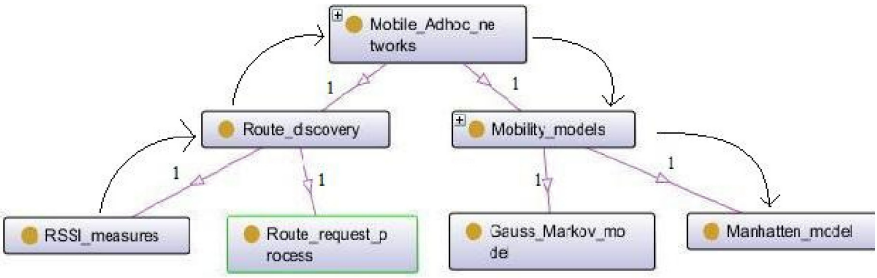

Fig. 1 Mapping snapshot to demonstrate hopping distance

3.4 Novelty Score Estimation

Each concept found in a document is assigned weight based on the number of other concepts which are present under the same category. For example, in a document, if there are n concepts associated with a concept X or those that come under the same category as concept X, the weight assigned to the concept X will be n+1. The summation of all the concepts weights gives the total value of the document. Consider there are n concepts that have been spoken in a document. The total weight W(d) of the document is given by

$$W(d) = \sum_{C_i}^{n} W(C_i) \quad (2)$$

Where $W(C_i)$ is the weight of the concept C_i. After the concepts in each document are weighted, then those documents are compared with the input document to compute the novelty score and those concepts that contributed to the novelty score. When a concept of the input document is found in the comparing document and if its weight is the minimum weight or equal to one, then that concept is ignored and its weight is deducted. But if its weight is more than the minimum weight, then its related or associated concepts are considered and are checked for their occurrence in the comparing document. If found, the number of those concepts found in the comparing document are used to deduct the weight of the comparing concept. This is repeated for all the concepts in the input document. The ratio of the new weight of the document to the initial weight of the document gives the novelty score. Consider $W(d|d_t)$ is the total weight of the computed concepts in the document d with respect to document d_t.

$$\text{Novelty score}(d|d_t) = W(d|d_t)/W(d) \tag{3}$$

Where W(d) is the initial weight of the document. The weights of concepts that are below or equal to the minimum weight are ignored and those that are above the minimum weight are considered to be the combination of concepts that contributed to the novelty score. These concepts are considered to be the novelty regions of the input document with respect to the comparing document. Once the concepts that contributed to the novelty of the document have been found, their term occurrence is searched in the document. The sections or paragraphs that contain those terms are retrieved to summarize the novelty regions of the document.

4 Results and Observation

Initial experiments were done to find the change in the novelty scores by introducing a survey paper in the corpus. We considered a survey paper of wireless sensor networks domain whose concepts matched the concepts defined in the Ontology at 7.2%. It is introduced in the dataset to observe if the novelty scores of each document changes. Similarly we observed the changes with respect to another survey paper which matched at 11.4% in the Ontology.

From the table 1, we observe that when a survey paper is introduced, the number of documents for which the novelty has been reduced depends on the number of concepts in the survey paper that are matched to the concepts defined in ontology. The number of novelty reduced documents when compared with survey pa-per(wireless sensor networks) matching 7.2% of concepts mapped in Ontology is less than that of when compared with survey paper(wireless sensor networks) matching 11.4% of concepts mapped in Ontology.

The table 2 shows the variation of novelty score of 10 research papers in the wire-less sensor networks domain. From the table 2, we can identify 2 cases which ex-plain the impact of survey papers on the research papers of that domain.

Table 1 Novelty reduced Documents in each domain

Sub-Domain Name	Number of Documents	Number of novelty reduced documents after comparison / Overall reduction in the Novelty (%)			
		With Survey Paper 1		With Survey Paper 2	
Wireless networks	66	24	4.8%	40	9.2%
Wireless Sensor	54	28	7.18%	34	12.79%
Adhoc	57	25	5.8%	39	14.3%
Multimedia	28	12	3.8%	16	13.1%
Peer to Peer	45	28	6.2%	26	7.8%
Cloud Computing	8	4	0.73%	3	3.9%
Network Traffic	82	30	4.72%	38	6.59%
Security	94	25	5.26%	41	5.42%
QOS	64	29	6.2%	27	10.7%
Mobile	31	13	4.81%	15	12.77%
Web service	33	14	3.03%	14	3.12%
Optical Network	57	21	5.52%	31	7.85%
Biometrics	19	4	6.34%	6	12.11%
Others	247	86	3.9%	96	5.68%

Table 2 Novelty scores for sample documents in Wireless Sensor Networks domain

Document ID	Novelty Score		
	Initial	Addition of Survey Paper 1	Addition of Survey Paper 2
1	0.86017	0.779661	0.533898
2	0.527851	0.527851	0.527851
3	0.652174	0.304348	0.304348
4	0.608374	0.490148	0.608374
5	0.426667	0.426667	0.266667
6	0.875	0.791667	0.441667
7	0.974217	0.961326	0.810313
8	0.86017	0.779661	0.533898
9	0.757895	0.610526	0.6
10	0.86017	0.779661	0.533898

1. The novelty score of the research paper reduces after comparing with survey papers when more concepts or concepts with more depth (more associated concepts) are matched with survey papers than any others.
2. The novelty score of the research papers does not reduce after comparing with survey papers when less number of concepts or concepts with less depth(less associated concepts) are matched with survey papers.

5 Conclusion and Future Work

On the basis of the studies, it is concluded that the system has shown promising results in identifying the relevant documents and filtering redundant documents with respect to a given document, using LDA and Kullback Leibler Divergence. The use of Ontology(Knowledge base) of a specific domain has enabled the system to measure the novelty of a document belonging to that particular domain and has highly contributed in the proper analysis of each document which in turn effects the overall performance of the system. Since measuring the novelty of a document is domain specific, the system requires a well-defined ontology for that particular domain. Thus the novelty present in a research paper of a specific domain is measured with respect to top relevant documents and given as a novelty score for the research paper. The concepts in the document that contributed to the novelty score are summarized to the user.

Our definition of novelty in research publication is based on the contribution of new combination of concepts and its depth. It can be further enhanced by analysing the contribution of each concept at the sentence level. The sentence level approach analyse the definition of concepts explored at deeper level which includes part of speech tagging, identifying entities etc. By improving the definition of concepts in the Ontology, the results can be enhanced.

Acknowledgments This work was funded by the Center for Technology Development and Transfer, Anna University Chennai under Approval no. CTDT-1/2360/RSS/2011 for Innovative project by Young faculty members under research support scheme.

References

[1] Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level, pp. 314–321. ACM (2003), http://doi.acm.org/10.1145/860435.860493

[2] Croft, X.L.W.: Answer updating approach to novelty detection. ACM (2004), http://dx.doi.org/10.3115/1220575.1220665

[3] Fernandez, R.T.: The effect of smoothing in language models for novelty detection. In: Future Directions in Information Access, FDIA 2007 (2007), http://www-gsi.dec.usc.es/_dlosada/fdia07.pdf

[4] Gamon, M.: Graph-based text representation for novelty detection, pp. 17–24. ACM (2006), http://dl.acm.org/citation.cfm?id=1654758.1654762

[5] Kok Wah Ng, L.C., Tsai, F.S., Goh, K.C.: Novelty detection for text documents using named entity recognition. IEEE (2007)
[6] Schiffman, B., McKeown, K.R.: Context and learning in novelty detection, pp. 716–723. ACM (2005), http://dx.doi.org/10.3115/1220575.1220665
[7] Stokes, N., Carthy, J.: First story detection using a composite document representation, pp. 1–8. ACM (2001), http://dx.doi.org/10.3115/1072133.1072182
[8] Tsai, F.S., Zhang, Y.: D2S: Document-to-sentence framework for novelty detection, vol. 29(2), pp. 419–433. ACM (November 2011), http://dx.doi.org/10.1007/s10115-010-0372-2
[9] Tsai, M.-F., Feng Tsai, M., Hsi Chen, H.: Some similarity computation methods in novelty detection. NIST (2002)
[10] Yang, Y., Zhang, J., Carbonell, J., Jin, C.: Topic-conditioned novelty detection, pp. 688–693. ACM (2002), http://doi.acm.org/10.1145/775047.775150
[11] Zhang, J., Ghahramani, Z.: A probabilistic model for online document clustering with application to novelty detection. In: Neural Information Processing Systems (NIPS), vol. 17 (2004)
[12] Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering, pp. 81–88. ACM (2002), http://doi.acm.org/10.1145/564376.564393
[13] Zhao, L., Zhang, M., Ma, S.: The nature of novelty detection, vol. 9(5), pp. 521–541. ACM (November 2006), http://dx.doi.org/10.1007/s10791-006-9000-x