

Building Domain Keywords Using Cognitive Based Sentences Framework

Zheng Xu, Weidong Liu, Yiwei Zhu and Shunxiang Zhang

Abstract As the novel web social media emerges on the web, large scale unordered sentences are springing up in the forms: news headlines, microblogs, comments and so on. Domain keywords extraction is very important for information extraction, information retrieval, classification, clustering, topic detection and tracking, and so on. Although these massive sentences contain rich information, their loose semantic association and highly unordered semantic organization make web users extremely difficult to capture the rich information due to the lack of semantic coherence. Sentence ordering is a significant research area focusing on obtaining coherent sentence orders which could assist web user to easily understand these unordered sentences. TextRank is a common graph-based algorithm for keywords extraction. For TextRank, only edge weights are taken into account. We proposed a new text ranking formula that takes into account both edge and node weights of words, named F2N-Rank. The results show our model can obtain coherent sentence orders with higher accuracy in less iterations. The proposed sentence ordering model can be applied in automatic text organization and summarization.

Keywords Domain keywords • TextRank • Sentence ordering

Z. Xu (✉)

The Third Research Institute of the Ministry of Public Security, Shanghai, China
e-mail: xuzheng@shu.edu.cn

W. Liu

Shanghai University, Shanghai, China

Y. Zhu

Zhejiang Business Technology Institute, Ningbo, China

S. Zhang

Anhui University of Science and Technology, Huainan, China

© Springer Nature Singapore Pte Ltd. 2018

N.Y. Yen and J.C. Hung (eds.), *Frontier Computing*, Lecture Notes
in Electrical Engineering 422, DOI 10.1007/978-981-10-3187-8_2

1 Introduction

Domain keywords can serve as a highly condensed summary for a domain, and they can be used as labels for a domain. Domain keywords should be ordered by the “importance” of keywords. With the boom of microblogs, massive unordered sentences are emerging on the web as a main message passing form. Although these sentences contain much useful information, loose semantic association and unordered sentence organization make web users lost in the large scale data when they face these massive unordered sentences. Web users normally expect these sentences are well ordered according to their semantic coherence since coherent sentence orders can assist them to easily understanding the content of these sentences. However, such sentence ordering problem is burdensome computation even though the 10 sentence scale is small.

In the study of keywords extraction, supervised methods [1–3] always depend on the trained model and the domain it is trained on. And in unsupervised methods [4, 5], algorithms based on term frequency and based on graph are the most common methods. Algorithms based on term frequency such as TF, ATF, ATF*DF, ATF*DF are easy to realize but their precisions are not very high. Algorithms based on graph, such as TextRank [1], are more effective than algorithms based on term frequency for they take into account the relationships among words. To overcome the above limitations, we adopt markov random field as special case of association link network which have been widely used in many tasks from learning technologies to knowledge discovery. Compared with association link network, markov random field has stronger ability in representation and inference since it implies association relation distribution and can make inference on the distribution. More importantly, our markov random field incorporates three cognitive logical structures which respectively guide sentence ordering model to link different sentence conditioned on different cognitive structures. What is more, we develop sound cognitive mechanistic for fast sentence ordering such as decision making process on cognitive logical structure, keywords spreading process and sentence activation process working on markov random field for ordering sentences.

TextRank is a common graph-based algorithm for keywords extraction. For TextRank, only edge weights are taken into account. We proposed a new text ranking formula that takes into account both edge and node weights of words, named F2N-Rank. The results show our model can obtain coherent sentence orders with higher accuracy in less iterations. The proposed sentence ordering model can be applied in automatic text organization and summarization.

2 Problem Formulation

Markov random field is a basic undirected probabilistic graphic model with outstanding abilities in semantic representation and inference. Inspired the above outstanding abilities of Markov random field, we propose semantic Markov random

field which is built by the limited number of association relations in power serials presentation represent and can inference the whole distribution of association relation in sentences to be order. In this paper, we can regard semantic Markov random field as a special case of association link network, which is constructed by association relations under different cognitive structures.

Sentence ordering task can be regarded as restrictive writing, since the content of sentences is known and the sentence order is unknown. The human beings’ task is to order these sentences for well semantic coherence. Based on cognitive process of writing [6–8], writing can be characterized as a “journey of discovery”. On this journey, association knowledge is continually activated and spreads to generate a coherent sentence order. Spreading activation model assumes that specific key-words distribute on semantic link network [9–13] and spreading activation process is a semantic processing on semantic link network, where keywords are continually spreading their influences into relevant keywords and these keywords can activate relevant sentence. Figure 1 shows sentence ordering process in restrictive writing, which spans on three memory modules: (1) long term memory, (2) short term memory and (3) working memory as follows.

- (1) Long term memory contains the relatively stable entities and relations. These entities and relations are stored by semantic link network. Besides, some cognitive logical structures are stored in this module. Keywords can spread towards different directions on semantic link network under different cognitive logical structures.
- (2) Short term memory contains cognitive logical structure schema and a word activation window with m activated keywords since Millers Law pointed out that the number of objects an average human can hold in short term memory is seven, plus or minus two. These keywords are activated from semantic link network by spreading activation process. Which keywords to be activated are conditioned on three different cognitive logical structures.

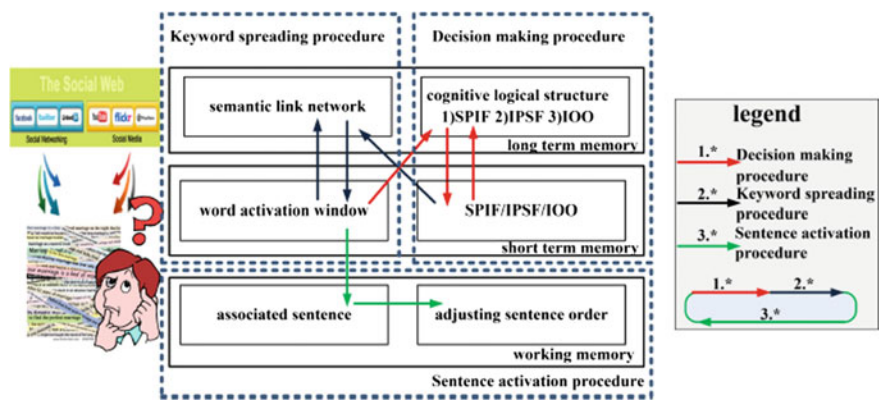


Fig. 1 The sentence ordering process

- (3) Working memory contains a newly generated sentence closely associated with activated keywords and a sentence order to be adjusted by spreading activation process. The spreading activation process will link all the unordered sentences toward well semantic coherence.

3 The Proposed Algorithm

TextRank algorithm only focuses on the relationship among nodes, and node weights are not taken into account. Equation (1) integrates TextRank formula with the node weight ($F(V_i)$).

$$FS(V_i) = (1 - d) * F(V_i) + d * F(V_i) * \sum_{V_j \in \text{In}(V_i)} \frac{w_{ji}}{\sum_{V_k \in \text{Out}(V_j)} w_{jk}} FS(V_j) \quad (1)$$

There are several formulas can be used to calculate the value of $F(V_i)$, such as TF, ATF, ATF*DF. ATF*DF is the most suitable of the three formulas because it takes into account both term frequency and document frequency. However, the simple combination of ATF and DF does not account for their proportions. Here, the idea of F-measure is introduced for calculating $F(V_i)$. The formulas are given as followings:

$$F(V_i) = \frac{(1 + \beta^2) * ATF(V_i) * DF(V_i)}{\beta^2 * ATF(V_i) + DF(V_i)} \quad (\beta = 2) \quad (2)$$

$$ATF(V_i) = \frac{\sum_{|D|} \frac{n_{i,j}}{\sum_k n_{k,j}}}{|\{d: t_i \in d\}|} \quad (3)$$

$$DF(V_i) = \log \frac{|\{d: t_i \in d\}|}{|D|} \quad (4)$$

The main steps of extracting domain keywords using F2N-Rank algorithm are as followings (Fig. 2):

- Step 1 Identify words (nouns, adjectives, and so on) that suitable for the task, and add them as nodes in the graph.
- Step 2 Identify relations that connect such words, and use these relations to draw edges between nodes in the graph. Edges can be directed or undirected, weighted or unweighted.
- Step 3 Calculate the weight of nodes in the graph.
- Step 4 Iterate the graph-based ranking algorithm until convergence.
- Step 5 Sort nodes based on their final score. Top T words are the domain keywords.

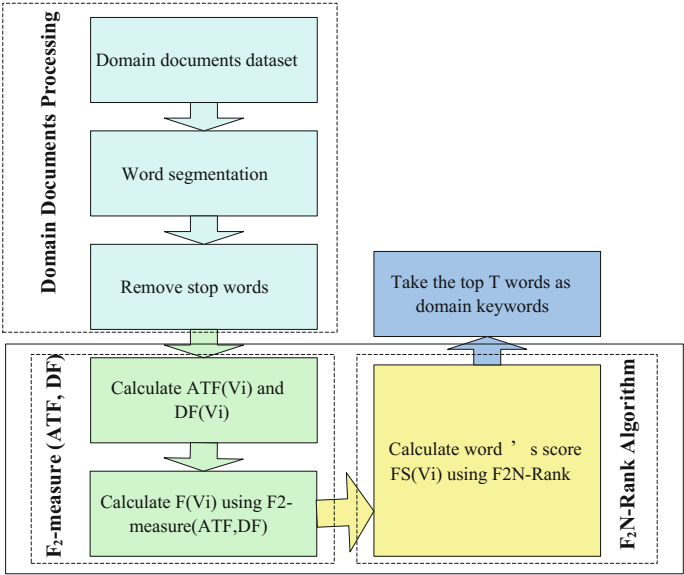


Fig. 2 The flow chat

Cognitive decision making is a core module of cognitive memory-inspired sentence ordering model (CM-SOM) since: (1) cognitive decision making module decides which cognitive logical structure and (2) different cognitive logical structures decides different sentence orders. As such, how to learn an optimal policy for cognitive logical structure decision making is an important issue for the cognitive decision learning model since such policy decides how to the logical structures shift during sentence ordering. To build the above ordering model and obtain a coherent sentence order, we propose the modules as follow.

- (1) Cognitive logical structure learning module. This module proposes three logical structure and construct cognitive logical structure based markov random field. Different cognitive logical structures and corresponding markov random filed shift in sentence ordering procedure;
- (2) Spreading and activation computation module. When a cognitive logical structure is selected, this module mainly makes keyword spreading and sentence activation based on makov random field, which is similar to the semantic association ability in human memory.
- (3) Cognitive decision making learning module. This module learns the decision making policy for shifting cognitive logical structure and such policy can guide which cognitive logical structure to develop sentence before a new sentence is linked.

4 Experiment and Results

We collect 2 datasets which consist of Reuters news as dataset 1 and paper abstracts as dataset 2. Domain data is used to learn logical structure and construct semantic markov random fields under different logical structures; training data is used to learn the decision making policy of cogitative logical structures. Test data is used to test the sentence ordering results generated by sentence ordering model. For each dataset, we will randomly select 50% texts as domain knowledge; 25% as texts as training data; 25% texts as test data from each dataset.

Dataset 1 includes 60,000 pieces of news for which each piece of these news has average 15.67 words and these news are crawled from Reuters website from March 2009 to August 2009. These news are about three domains including health, environment, internet. Dataset 2 includes 50,000 paper abstracts for which each one has average 13.27 words and these papers are from Association for Computing Machinery-digital Library. These papers cover 10 different categories including data mining, machine learning, algorithm and so on.

To evaluate the performance of ranking Tibetan religious keywords, we conducted a performance measurement using precision. Now, we discuss the evaluation of three different ranking algorithms. We compared algorithms which are: F2N-Rank, TextRank and ATF*DF. Results are shown in Fig. 3 by measuring the precision for top N keywords. We can see that F2N-Rank clearly outperformed both TextRank and ATF*DF. For F2N-Rank, TextRank and ATF*DF, the average precision are 78.6, 62.2 and 49.2%. The improvement over TextRank is around 16% in average precision and 29% over ATF*DF. Using F2N-Rank for domain keywords extraction has showed better results.

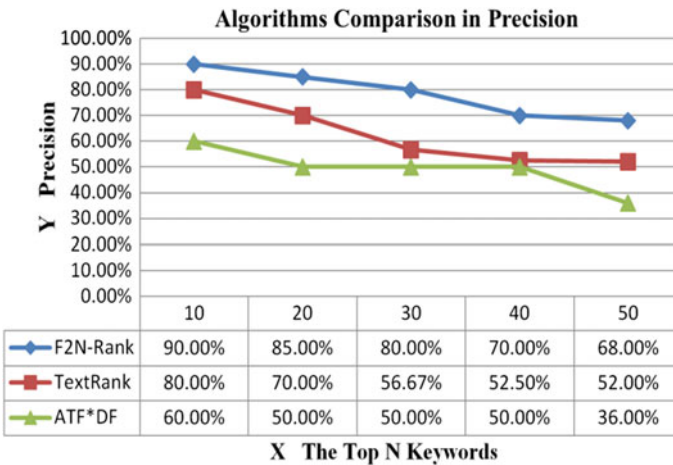


Fig. 3 Algorithm Comparison in Precision

5 Conclusions

Domain Keywords extraction is important for many applications of Natural Language Processing. TextRank is a common graph-based algorithm for keywords extraction. For TextRank, only edge weights are taken into account. We proposed a new text ranking formula that takes into account both edge and node weights of words, named F2N-Rank. The results show our model can obtain coherent sentence orders with higher accuracy in less iterations. The proposed sentence ordering model can be applied in automatic text organization and summarization.

Acknowledgements This work was supported in part by the National Science and Technology Major Project under Grant 2013ZX01033002-003, in part by the National High Technology Research and Development Program of China (863 Program) under Grant 2013AA014603, in part by the National Science Foundation of China under Grant 61300202, in part by the China Post-doctoral Science Foundation under Grant 2014M560085, and in part by the Science Foundation of Shanghai under Grant 13ZR1452900.

References

1. Mihalcea R, Tarau P, "TextRank, Bringing order into texts", Association for Computational Linguistics, 2004.
2. Frank, Eibe, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning, "Domain-specific keyphrase extraction", In Proceedings of the 16th International Joint Conference on Artificial Intelligence, pp. 668–673, 1999.
3. Medelyan, Olena, Eibe Frank, and Ian H. Witten, "Human-competitive tagging using automatic keyphrase extraction", In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1318–1327, 2009.
4. Tomokiyo, Takashi and Matthew Hurst, "A language model approach to key phrase extraction", In Proceedings of the ACL Workshop on Multiword Expressions, 2003.
5. Turney, Peter, "Learning algorithms for key phrase extraction. Information Retrieval", Vol. 2, pp. 303–336, 2000.
6. L. Flower, J. R. Hayes, A cognitive process theory of writing, *College composition and communication* (1981) 365–387.
7. A. C. Graesser, M. Singer, T. Trabasso, Constructing inferences during narrative text comprehension., *Psychological review* 101(3) (1994) 371.
8. X. Luo, J. Zhang, F. Ye, P. Wang, C. Cai, Power series representation model of text knowledge based on human concept learning, *Systems, Man, and Cybernetics: Systems*, IEEE Transactions on 44(1) (2014) 86–102.
9. C. Hu, Z. Xu, et al. Semantic Link Network based Model for Organizing Multimedia Big Data. *IEEE Transactions Emerging Topics in Computing*, 2(3): 376–387 (2014).
10. X. Luo, Z. Xu, J. Yu, and X. Chen. Building Association Link Network for Semantic Link on Web Resources. *IEEE transactions on automation science and engineering*, 8(3):482–494, 2011.
11. Z. Xu, et al. Knowle: a Semantic Link Network based System for Organizing Large Scale Online News Events. *Future Generation Comp. Syst.* 43–44: 40–50 (2015).
12. Z. Xu et al. Incremental building association link network. *Computer systems science and engineering*, 26(3):153–162, 2011.
13. Zheng Xu, Xiangfeng Luo, Wenjun Lu. Association Link Network: An Incremental Semantic Data Model on Organizing Web Resources. *ICPADS 2009*: 793–798.

Frontier Computing

Theory, Technologies and Applications FC 2016

Yen, N.Y.; Hung, J.C. (Eds.)

2018, XIX, 1011 p. 377 illus., 248 illus. in color.,

Hardcover

ISBN: 978-981-10-3186-1