Lecture Notes in Computer Science:
Multi-Word Wrapping Method for Enhanced Performance of Neural Network Algorithm in Big Data Analysis

Su-Kyoung Kim[1], Kee-Hong Ahn[1], Jin-Woo Lee[1], and Yoon-Sung Jo[1]

[1]125, Dongseo-daero, Yuseong-gu, Daejeon, 305-719
National Hanbat University, Korea
{kimsk, khanh}@hanbat.ac.kr, ljw0084@nate.com, mm1023@naver.com

Abstract. This study suggests the way of documents recommend in neural algorithm based on multi-word wrappers for high quality output in big data environment. Neural algorithm has advantage of prediction after a lot of learning. Neural algorithm has good predicting rate, generation of big data maybe can be good environment for learning. But many data can include worthless information. This approach can't lead to efficient for documents recommend. So we suggest multi-word wrapper for enhancing result. Our model use both dictionary perspective and user perspective. We expect our suggesting model result that use multi-word wrapper has some advantage than general approach.

Keywords: Big-data, Neural Network Algorithm, Machine Learning, WordNet, Semantic Processing, Document Recommendation

1   Introduction

Recently many studies to retrieve document user want have newly focused as the advent of big-data, which is generated in real time. The way of document recommendation in this environment is important. Most people want useful information. But there are a lot of disadvantage in terms of obtaining useful information due to tremendous data volume. For such reason, the need of technology of personal document recommendation has been increasing. Existing general model such as TF-IDF, cosine similarity and Jaccard similarity has limit to identify characteristic of user. To solve this problem, we suggest multi-word-wrapper (MWW) that can reflect user perspective. MWW consist of two levels. First level measure value of word based on document topic and two level processes based on user information. In addition we get relations between words using WordNet. It can obtain more relational result with topic. Then input data about neural algorithm will be good.

2 Related Research

2.1 Neural Algorithm

Representative learning algorithms are genetic algorithm and neural algorithm that try move human study process to computer. Neural algorithm organize network through many neural cell consist of neural model based on nerve cell.

$$u = \sum_i x_i w_i - v$$

(1)

Neural consist of many input data, weight of input data and critical point. We can get output through various value that minus from critical value. If we obtain 'u', it use next formal. After getting value 'u', it apply appropriate function. Then we can get 'z'. Appropriate function has two. Step function has binary output. Sigmoid function has continuous value that 1 to 0. We use sigmoid function. Output of sigmoid function can use judgment to document similarity.

$$f(u) = \frac{1}{1 + e^{-u}}$$

(2)

2.2 WordNet Analysis

This paper was try to add the relation between words and to process mean of word using structure of WordNet. Because we apply two perspective that documents perspective and users for value of word. Basically WordNet consist of linked list and each node has synsets relations about one word. Next tables are sample of classification about noun and verb.

Table 1. Noun Classification Structure

Classification of Noun		
{act, action}	{animal, fauna}	{artifact}
{attribute}	{body, corpus}	{cognition}
{communication}	{event}	{feeling}
{food}	{group}	{location, place}
{motive}	{natural object}	{natural}

{person}	{plant, flora}	{possession}
{process}	{quantity}	{relation}
{shape}	{state, condition}	Substance}
{time}		

Table 2. Verb classification structure

Classification of Verb		
{bodily care}	{change}	{cognition}
{communication}	{competition}	{consumption}
{contact}	{creation}	{emotion}
{motion}	{perception}	{possession}

Synset is set of similar word. Fig. 1 is a example in relation network of WordNet. That include hypernymy, hyponymy and meronymy. WordNet has many relations between nodes. Node has four part of speech (Noun, Verb, Adjective, Adverb). Words of WordNet have many relations for hypernym, hyponym, antonym and meronym etc.

3   Suggested Method

3.1   Multi-Word Wrapper Structure

We choose machine learning mechanism to accurate measure in document similarity. Base is mechanism is MWW. It is quantize to value of keyword in document through two wrapping. Extracting keyword in document is based on document characteristic. First extracted word is measured by document topic. And apply first wrapping function. Continuously previous value applied, second wrapping function. It is process run by user information. In other wise, if correspond word tendency and user tendency, word value is increase. The result, we can extract keyword that has user perspective and can calculate of similarity using extracted keyword. Fig. 2 is suggesting diagram.
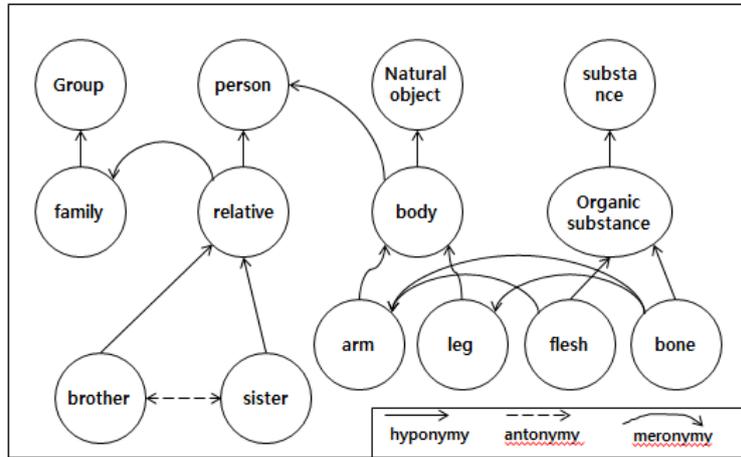
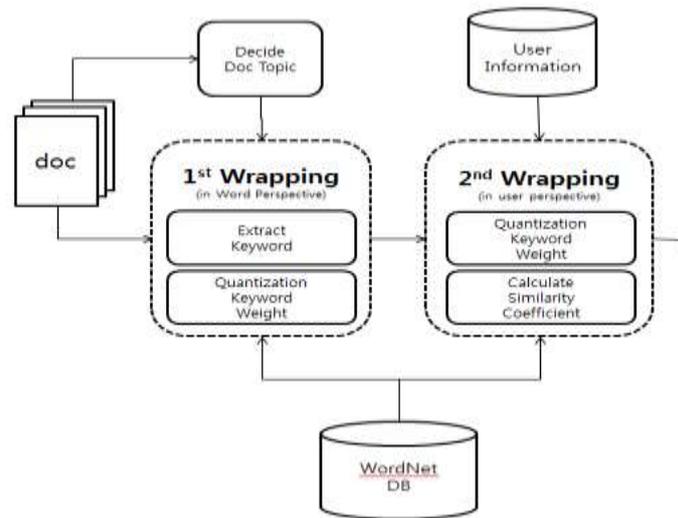**Fig. 1** Diagram of WordNet Classification.

Fig. 2. Concept Diagram of Multi-Word Wrapper for Neural Network Algorithm

3.2   Preprocessing of Sentence using WordNet

A Word in document is measured by topic keyword. Classification of WordNet follow table 1 and table 2. We add the relation to word that "Antonym", "Hypernym", "Hyponym", "Holonym" and "Meronym". Table 3 show the sentence that is "I am

happy becase Shellfish is very delicious", as you can see, sentence separated token, classification and description using API supplied by WordNet.

Table 3. Example of Sentence Parsing using API of WordNet

Setence		I am happy becase Shellfish is very delicious
Token	Classification	Description
I		
am	[be]	have the quality of being (copula, used with an adjective or a predicate noun)
happy	[felicitous,happy]	marked by good fortune
because		
Shellfish	[shellfish]	meat of edible aquatic invertebrate with a shell
is	[be]	have the quality of being (copula, used with an adjective or a predicate noun)
very	[very,really,real,rattling]	used as intensifiers; 'real' is sometimes used informally for 'really'; 'rattling' is informal
delicious	[Delicious]	variety of sweet eating apples

3.3 Whole Function Structure of Our Processing

We input many document that already similar. Each document have various information which many kinds of domain and semantics. For example topic, date, site address and author information etc. We decide title using this information. First wrapping, we allocate value to word of extracted frequency of word in document based on document. And it hand over to second wrapper. Second wrapping is same process but it is little different. Second our wrapper allocating the value to word based on user information. In this way, various perspectives have meaning for one word. After word wrapping is sorted by descending order. And we compare document for similarity using the top word that ranked from 5 to 10. Each word in each document is enough to be compare object, because it word properly reflecting topic in document and user information. We use a "Pearson Correlation Coefficient (PCC)" for relation between documents. It has two relations that positive correlation, negative correlation. And use shingling for judgment whether compound word or simple word. Finally we use TFiDF which word has best value in document. After this process, we can get numerical words. So this word to be new input data for neural algorithm. And this process repeat that each time make weight little differ until that algorithm say 'same'. This process can find best similarity weight.
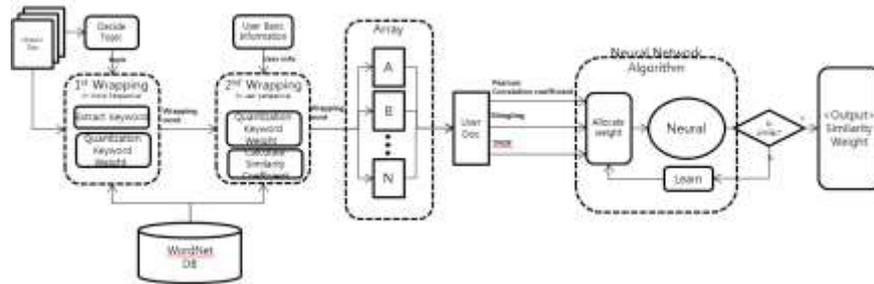
Fig. 3. Whole Function Structure for Processing of Our Method

4   Conclusion

We has many attempting for variety raise similarity analysis result base on neural algorithm. We begging word of the qualitative relationship between the word. In succession, we add variety elements that word perspective and user perspective. Finally, although we want verification this process system that combination of different formulas, this system construct and test not yet. Suggested idea in this study, since basic status for data analysis, design and implementation is not yet complete. So, we more contribute and confirm our idea, next we design the detail function and implement the test system, and will show excellent of our idea.

References

1. http://wordnet.princeton.edu/wordnet/documentation/
2.   National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov
3. Yun Her, Su-Kyoung Kim, Young-Taek Jin, "A Context-Aware Framework using Ontology for Smart Phone Platform," International Journal of Digital Content Technology and its Applications, vol.4, no.5, pp.159--167, Aug (2010)
4. S.Kim, K.Hong, "A Study of Dynamic Web Ontol ogy for Comparison- shopping Agent based on Semantic Web", Journal of Intelligence and Information Systems, Korea Intelligent Information System Society Quarterly, pISSN : 1229-4152, v.11, no.2, pp. 31-- 45 (2005)
5. S. Kim, " Using Description Logic and Rule Language for Web Ontology Modeling", Proceedings of Korea Intelligent Information System Society (KIISS) spring conference, pp.277-- 285 (2007)

6. Kim-su kyoung, Kim-gun woo,"customized searching service which considerate the user's intention based on ontology." Intelligence information research, v.18, no.4, pp.19--28, December (2012)
7. Lee-jin woo, Kim-su kyoung, "Analysis and Supplement Research of Hand Model", Korea Science Computer Information Society (KSCI review), Vol.20 No.2 (2012)
8. Su-kyoung Kim, Yoon-sung Ch, Ki-Hong Ahn, "Mobile Search Interaction Framework based on User Experience Reusing", JNIT: Journal of Next Generation Information Technology, Vol. 4, No. 2, pp. 20--28 (2013)
9. Sukyoung Kim, Kihong Ahn: User-intent custom adaptive search service based on ontology and similar-word retrieval algorithm composition. ICITST pp.121--126, December (2012)
10. Jurgen Angele, Michael Kifer, and Georg Lausen, "Ontologies in F-Logic", International Handbooks on Information Systems, DOI 10.1007/978-3-540-92673-3, pp.45--70 (2009)
11. Jeff Z. Pan, "Resource Description Framework", International Handbooks on Information Systems, DOI 10.1007/978-3-540-92673-3, pp.71--90 (2009)
12. Pascal Hitzler and Bijan Parsia, "Ontologies and Rules", International Handbooks on Information Systems, DOI 10.1007/978-3-540-92673-3, pp. 111--132 (2009)
13. Bo-Hyun Yun, Dea-Ho Back, Sang-Bum Kim, Kyung-Soo Han, Hea-Chang Lim, An Information Retrieval Method of Alleviating Semantic Term Mismatches with Interactive Query Expansion, the Korea Institute of Information Science, (1999)
14. Chantal Taconet, Zakia Kazi-Aoui, "Context-awareness and Model Driven Engineering", E-Commerce application scenario. ICDIM (2008)
15. MobiLife, http://www.ist-mobilife.org, (2008)
16. Apache Jena http://jena.apache.org/