A Semantic Analysis Method for Concept Map-based Knowledge Modeling

Jin-Xing Hao[1] Angela Yan Yu[2] Ron Chi-Wai Kwok[2]

[1]School of Economics and Management, Beihang University, Beijing 100191, China
[2]Department of Information Systems, City University of Hong Kong, Hong Kong, China

Abstract

Semantic analysis of a concept map plays an important role in translating human knowledge in the form of concept maps into rigorous and unambiguous representations for further processing by computers. However, recent research limits in the literal analysis of concept labels and concept relatedness that is derived from the structure of concept maps. In this study, we propose and evaluate a semantic analysis method which incorporates a formal representation of a concept map and WordNet-based algorithms to compute semantic similarity. As a fundamental element of knowledge modeling, the work presented in the study implies important contributions in business intelligence research and practice.

Keywords: Concept map, semantic analysis, knowledge modeling

1. Introduction

Knowledge modeling is the process of representing combinations of data or information into a reusable format for the purpose of preserving, analyzing, and sharing. It is a critical element and a prerequisite for reaching true business intelligence [1, 2]. Classical knowledge modeling methodologies, e.g., frame system and descriptive logics, following the tradition initiated by John McCarthy [3], are non-ambiguous and straightforward to computer but present a technical barrier for human who are unfamiliar with these formalisms. Concept map based knowledge modeling has marked a main trend to represent knowledge from computer-centered to human-centered [4]. It addresses the importance of creating knowledge bases that are natural to share and process by people rather than by software systems.

According to Novak and Gowin [5] and Trochim [6], a concept map is a diagram indicating inter-relationships among concepts and representing conceptual frameworks within a specific knowledge domain. Concept mapping has been widely accepted as a knowledge elicitation tool for its flexibility and ease-of-use to represent human's mental model. But when concept mapping is used as a knowledge modeling method, it faces challenges on the mediation mechanisms to translate elicited human knowledge in the form of concept maps into rigorous and unambiguous representations for further processing by computers. Structural analysis and semantic analysis of a concept map are two important aspects to achieve the above mechanisms [7].

Structural analysis of a concept map focuses on its structure which reflects how the concepts within a concept map are inter-related. Concepts are abstracted as vertices, and relationships among them are as arcs. Analysis and inference can therefore be conducted by using the generated graph structure. In contrast,

semantic analysis of a concept map captures the meaning of concepts and their relations. Traditionally, the semantic analysis is mainly conducted by human experts. The process is not only labor-intensive and time-consuming but also inevitably involves human bias.

In our study, we propose an automatic semantic analysis method for concept map based knowledge modeling. Specifically, we examine the lexical composition of concepts and propose a formal representation of concepts. Then we design WordNet-based algorithms to compute the concepts' similarity. Further we evaluate and optimize the parameters using a set of carefully designed experiments. As a fundamental element of knowledge modeling, the work presented in the study has wide application in business intelligence research and practice.

The following sections are as follows: In section 2, we review the related work of semantic analysis of concept map. In section 3, we propose the formal representation of a concept map. In section 4, we elaborate the major algorithms of our analysis method. In section 5, we evaluate these algorithms using a set of experiments. Finally, we conclude the paper and point out the limitations and future works of the study.

2. Related Work

Concept maps are rooted in Ausubel's Assimilation theory [8]. Decades of research and practice has demonstrated that concept maps can aid people of different ages to examine many fields of knowledge [9]. They offer the flexibility of natural language and have the advantage of inducing their creators to organize their knowledge in a structured fashion, where concepts and their connections can be directly recognized.

Concept maps, as illustrated in Fig. 1, are composed of nodes (or ovals) that represent concepts and links (or arcs) that connect nodes to represent the relationships between concepts.

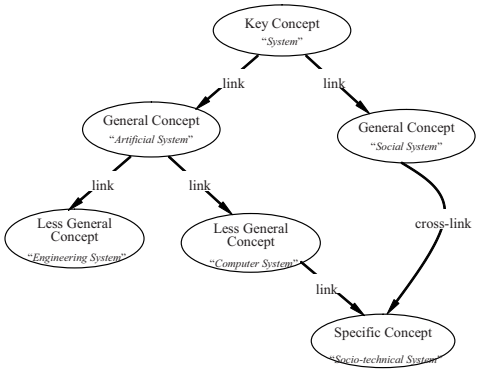

Fig. 1: A sample concept map.

Each node-link-node triplet forms a proposition with a meaningful statement about the object or event. These nodes and links are arranged in a hierarchical fashion with the most general concept at the top followed by more and more specific ones. The segments of a concept hierarchy represent different knowledge domains within the concept map.

Formally speaking, the topological structure of a concept map is a digraph. Many studies on analyzing concept similarity are based on this formalism, such as the proposition based method [10], the topological dimensions based method [11], the vector space model based method [12], and the method based on the similarity and dissimilarity of description logics [13]. Each method needs a certain way to represent concepts. For example, the proposition based method represents the concepts by their labels with related propositions. The topological dimensions based method needs to consider the structural position

of a concept as weighting factor. The vector space model based method computes term-frequency vector with inverse-document frequency adjustment with proper term weights. And the description logic approach requires logic representations that are derived from the structure.

However, these research on concept analysis of concept map limits in the literal analysis of concept labels and concept relatedness that is derived from the structure of a concept map. The true semantics of a concept is still absent from the analysis.

In concept maps, the representational vocabulary is non-standardized, and the link names are generic, therefore the most significant information source is usually on the concepts rather than on the links. In this study, the semantic analysis focuses on concepts in a concept map.

3. Formal Representation of a Concept Map

3.1. Conceptual foundations

We argue that there are two levels of concept labels in a concept map: simple label and complex label. A simple label refers to the terminology or a single word in one specific domain, while a complex label is the meaningful composite of simple labels to represent the advanced and abstract ideas. For example, "motivation" is a simple label, while "the motivation of sale representatives" is a complex label. The labels of concepts in a concept map are generally complex labels.

According to cognitive theories [14], simple labels, as a form of ontology, are relatively easy to establish the hierarchical relations among them. In contrast, it is difficult, if not impossible, to create the hierarchical relations among complex labels. A complex label is comprised of several simple labels. These simple labels work together to express the advanced and specific concepts within one domain and to reflect the higher levels of knowledge such as causal knowledge. Thus, it is difficult to articulate the parent or child class of complex labels.

Recent cognitive theories, such as connectionist school [14], advocate the network structure of human mental model. Therefore, it is preferred to establish the network relations among complex labels by their semantic similarities. The hierarchical relations among simple labels or ontology can be utilized to generate the lexical relatedness among complex labels, which works as the surrogate measure of their similarities.

Based on the above considerations, we propose a formal representation of a concept map.

3.2. Definition of a concept map

A concept map represents general knowledge on a domain. A concept map is a 3-tuple

$$S = (T_C, T_R, O) \quad (1)$$

where,

T_C , the set of concepts labels, is a complete graph to reflect their semantic relatedness;

T_R , is a the set of dyadic relations, T_C and T_R are disjoint;

O, is a ontology which consists of simple labels and their hierarchy in the specific knowledge domain.

4. WordNet-based Concept Semantic Similarity Computing Method

According to the formal representation, most of the concept labels in a concept map are complex labels which are comprised of simple labels. Therefore the semantic analysis of concepts is

comprised of two major components: simple label semantic similarity and complex label semantic similarity.

4.1. Simple label semantic similarity

Several upper ontology and domain ontologies can serve as foundations to compute simple label semantic similarity. In this study, we are interested in WordNet [15], a broad coverage lexical network of English words. What differentiates WordNet from other dictionaries is that the organization of words is not simply alphabetic, but is based on the senses (semantics) of the words. During the twenty years of development, WordNet has become one of the most widely adopted resources for semantic analysis.

In WordNet, nouns, verbs, adjectives, and adverbs are each organized into networks of synonym sets (synsets) that each represents one underlying lexical concept and are interlinked with a variety of relations. A polysemous word will appear in one synset for each of its senses. The noun network of WordNet was the first to be richly developed, and most researchers' work is limited to this part.

Three basic measures are widely used for computing two words semantic relatedness in our study:

1) The length of the shortest path in WordNet is from synset c_i to synset c_j (measured in edges or nodes is denoted by len(c_i,c_j).

2) We write lso(c_1,c_2) for the lowest super-ordinate (or most specific common subsumer) of node c_1 and c_2.

3) The Resnik's information content of two words,

$$\mathrm{IC}(w_1, w_2) = -\log p(\mathrm{lso}(w_1,w_2)), \quad p(c) = \frac{\sum_{w \in W(c)} \mathrm{count}(w)}{N} \tag{2}$$

where $W(c)$ is the set of words (nouns) in the corpus whose senses are subsumed by concept c, and N is the total number of word (noun) tokens in the corpus that are also present in WordNet. In Resnik's experiments, the corpus are the one-million-word Brown Corpus of American English [16].

Based on the basic measures, some advanced measures have been proposed to enhance the accuracy. Prior studies [17] have suggested that Jiang and Conrath's Combined Approach [18] and Li et al.'s Appraoch [19] have the best performance. Due to the fact that multiple information sources can be synthesized to improve the measure performance [19], we propose a combination formula to synthesize Jiang and Conrath's Combined Approach $dist_{JC}$ and Li et al.'s Appraoch sim_{LI} after a set of experiment elaborated in Section 5.

$$\mathrm{sim}_{JC\_LI}(c_1,c_2) = 0.3 \cdot \mathrm{dist}_{JC} + 0.7 \cdot \mathrm{sim}_{LI} \tag{3}$$

where

$$\begin{aligned} \mathrm{dist}_{JC}(c_1,c_2) &= \mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2 \times \mathrm{IC}(lso(c_1,c_2)) \\ &= 2\log p(lso(c_1,c_2)) - (\log p(c_1) + \log p(c_2)) \end{aligned} \tag{4}$$

and

$$\mathrm{sim}_{LI}(c_1,c_2) = e^{-0.2 \cdot l} \cdot \frac{e^{0.45 \cdot h} - e^{-0.45 \cdot h}}{e^{0.45 \cdot h} + e^{-0.45 \cdot h}}. \tag{5}$$

4.2. Complex label semantic similarity

Complex labels are comprised of simple labels. However, they are not complex enough to use traditional information retrieval methods to compute the similarity between them. Traditional information retrieval methods, such as space vector model, always contain

thousands of elements for one vector. In the concepts of causal maps, they generally have four to five words. Therefore, we must compute their similarity with the help of the simple concept similarity described above. The formula we use generally follows Li et al. [20],

$$S(T_1,T_2)=\delta S_s+(1-\delta)S_r=\delta\frac{\mathbf{s_1}\cdot\mathbf{s_2}}{\|\mathbf{s_1}\|\cdot\|\mathbf{s_2}\|}+(1-\delta)\frac{\|\mathbf{r_1}-\mathbf{r_2}\|}{\|\mathbf{r_1}+\mathbf{r_2}\|} \quad (6)$$

where $\delta \leq 1$ decides the relative contributions of S_s, semantic information, and S_r, word order information, to overall similarity computation. According to Li et al. [20], δ is suggested to 0.85.

5. Evaluation

5.1. Evaluation of simple label semantic similarity

The general strategy to evaluate these measures is to compare them with human judgments which clearly give the best assessment of the "goodness" of a measure. Its main drawback lies in the difficulty of obtaining a large set of reliable, subject-independent judgments for comparison. In our study, we employ the best benchmark experiment data to date.

In 1965, Rubenstein and Goodenough [21] obtained "synonymy judgments" from 51 human subjects on 65 pairs of words. The pairs ranged from "semantically unrelated" to "highly synonymous" on the scale of 0 to 4. Another commonly used set is from an experiment of Miller and Charles [22]. They chose 30 pairs from the original 65, taking 10 from the "high level (between 3 to 4), 10 from the intermediate level (between 1 and 3), and 10 from the low level (0 to 1) of semantic similarity", and then obtained similarly judgments from 38 subjects, given the same instructions as above.

In this study, we use both sets of words. We use Miller and Charles's (M&C) word set to fine tune and develop the similarity formula, and use the full Rubenstein and Goodenough's (R&G) word set to test the finalized formula.

Measures	Correlation (significance)
$Dist_{JC}$	0.8269 ($p < 0.0001$)
Sim_{LI}	0.8688 ($p < 0.0001$)
Sim_{JC_LI}	0.8735 ($p < 0.0001$)

Table 1: Correlation of measures with R&G's rating.

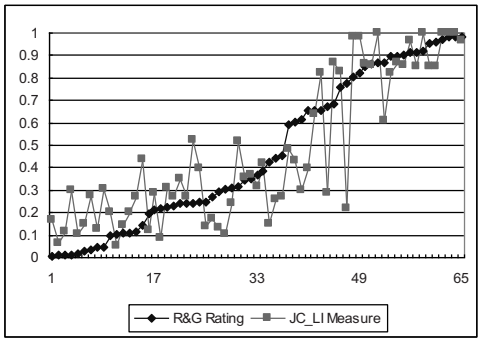

Fig. 2: Line graph of JC_LI measure with R&G rating.

Table 1 summaries the correlation of the candidate measures with R&G's human rating. We also show the line graph of JC_LI measure with R&G Rating as Fig. 2. The graph shows that the JC_LI similarity is quite close with the human rating regarding the upbound of human rating 0.8848 [17].

5.2. Evaluation of complex label semantic similarity

Similar to simple label semantic similarity, we need a benchmark to evaluate complex label semantic similarity. Following Li et al. [20], we

generate a complex label set by replacing the noun pairs in Rubenstein and Goodenough (R&G) word set with the definitions from the Collins Cobuild dictionary [23]. Cobuild dictionary definitions are selected because its vocabulary and grammatical structures are similar with the human daily languages. The dictionary is constructed using information from a large corpus, the Bank of English, which contains 400 million words. When there is more than one sense in the list, we choose the first noun sense. Thus, we can compare the complex label similarity measures with R&G's rating to determine the optimal thresholds. It is also noted that complex nodes have more information then single word. Human may perceive the similarity of the word pair differently when provided with definitions. Therefore, we also use the human rating of the sentence data set as shown in Li et al. [20] to adjust the parameters. The full 65 word definitions set is used as test data set.

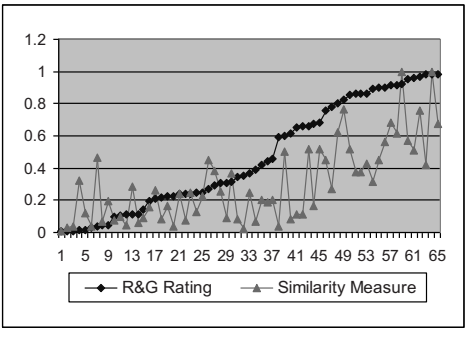

Fig. 3: Line graph of complex label similarity measure with R&G rating.

Fig. 3 illustrates the line graph of our similarity measure based on the 65 word definitions with R&G ratings. The correlation of them is as high as 0.72.

6. Conclusions and Future Works

In this study we elaborate our preliminary practice on proposing a semantic analysis method for concept map based knowledge modeling. It is just a meaningful attempt, although there are many limitations of the study which point to future studies. For example, we will improve the definition of a concept map based on Sowa's Conceptual Structure Theory [24]. We will try to propose new analyzing models by synthesizing both semantic and structural information of a concept map [11]. At the current stage, this study still implies theoretical contributions in improving mechanisms to transform knowledge models that are built by concept map technologies into the rigorous and unambiguous representations. At same time, it has practical applications in various business domains, such as knowledge assessment, strategy making, and collaborative decision making.

Acknowledgments

The work described in this paper is supported by the Fundamental Research Funds for the Central Universities of China (No. YWF-10-03-007) and the Teaching Development Grant of the City University of Hong Kong, (No. 6000177).

References

[1] R. J. Brachman and H. J. Levesque, *Readings in knowledge representation*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. , 1985.

[2] H. R. Nemati, D. M. Steiger, L. S. Iyer, and R. T. Herschel, "Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing," *Decision Support Systems,* vol. 33, pp. 143-161, 2002.

[3] J. McCarthy and P. Hayes, "Some philosophical problems from the standpoint of artificial intelligence,"

Standford Univ Calif., Dept of Computer Science 1968.

[4] A. G. Maguitman, "Intelligent support for knoweldge capture and construction," Indiana University, 2004.

[5] J. D. Novak and D. Gowin, *Learning How to Learn*. Cambridge: Cambridge University Press, 1984.

[6] W. Trochim, "An introduction to concept mapping for planning and evaluation," *Evaluation and Program Planning,* vol. 12, pp. 1-16, 1989.

[7] J.-X. Hao, R. C.-W. Kwok, R. Y.-K. Lau, and A. Y. Yu, "Predicting problem-solving performance with concept maps:An information-theoretic approach," *Decision Support Systems,* vol. 48, pp. 613-621, 2010.

[8] D. P. Ausubel, J. D. Novak, and H. Hanesian, *Educational psychology: A cognitive view*. New York: Holt, Rinehart and Winston, 1978.

[9] J. D. Novak, *Learning, Creating, and Using Knowledge: Concept Maps ™ as Facilitative Tools in Schools and Corporations*. Mahwah, NJ: Lawrence Erlbaum, 1998.

[10] Y. Zhu, R. Zhang, and I. Ahmad, "Applying Concept Similarity to the Evaluation of Common Understanding in Multidisciplinary Learning," *Journal of Computing in Civil Engineering,* vol. 24, pp. 335-344, 2010.

[11] D. B. Leake, A. Maguitman, and A. Canas, "Assessing Conceptual Similarity to Support Concept Mapping," in *Proceedings of FLAIRS-02*, 2002, pp. 168-172.

[12] D. B. Leake, A. Maguitman, and T. Reichherzer, "Topic Extraction and Extension to Support Concept Mapping," in *Proceedings of FLAIRS 2003*, 2003, pp. 325-329.

[13] C. d'Amato, N. Fanizzi, and F. Esposito, "A semantic dissimilarity measure for concept descriptions in ontological knowledge
bases," in *2nd Int. Workshop on Knowledge Discovery and Ontologies*, Porto, Portugal, 2005.

[14] J. R. Anderson and C. Lebiere, *The Atomic Components of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates, 1998.

[15] C. Fellbaum, "WordNet: An Electronic Lexical Database," Cambridge, MA: MIT Press, 1998.

[16] W. N. Francis and H. Kucera, *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin, 1982.

[17] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics,* vol. 32, pp. 13-47, 2006.

[18] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, Taiwan, 1997, pp. 19-33.

[19] Y. H. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity using multiple information sources," *IEEE Transactions on Knowledge and Data Engineering,* vol. 15, pp. 871-882, 2003.

[20] Y. H. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Transactions on Knowledge and Data Engineering,* vol. 18, pp. 1138-1150, 2006.

[21] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM,* vol. 8, pp. 627-633, 1965.

[22] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes,* vol. 6, pp. 1-28, 1991.

[23] J. Sinclair, "Collins COBUILD advanced learner's English Dictionary.," Glasgow, UK: HarperCollins Publishers Limited, 2006.

[24] J. F. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*. New York, NY: Addison-Wesley, 1984.