

On-Line Text Mining and Recommendation based on Ontology and Implied Sentiment Inclination

Hui Yang^{***}, Zhe Wang^{*}, Hao Xu^{*}

^{*} College of Computer Science and Technology Jilin University, Changchun 130012, China

^{**}College of Electronic Science and Engineering Jilin University, Changchun 130012, China

yanghui2012@jlu.edu.cn, wz2000@jlu.edu.cn, haoxu@jlu.edu.cn

Abstract—Most of the exist Web search engines utilize matching the query keywords to pieces of information approach to identify of the data satisfying user's request . These methods are not only inefficient, but also wasted a lot of user's time to find a satisfactory results. In order to improve the problem above, we presented a different approach to identify user's request that attracts more interest is called a ontology-based and implied sentiment search. Ontology is a tree structure which represent specifications of concepts and relations among them. Ontology play a central role in semantic web applications by providing a shared knowledge about the objects in real world. This paper proposes a method for determining semantic similarity between concepts and implied sentiment defined in ontology. Unlike method exist that use ontological definition of concepts for similarity assessment, the presented approach also focuses on the relations between concepts and their implied sentiment inclination. Our method is able to determine similarity not only at the definition level, but also is able to evaluate similarity of implied sentiment of information that are instances of concepts. In addition, the method allows for context-aware similarity assessment .Experimental comparison of our on-line text mining approach against other techniques known in the literature shows satisfying results.

Keywords—Sentiment Inclination Analysis, Text Mining, Ontology, similarity algorithm, recommendation method

I. INTRODUCTION

With the fast growing of Web content and available data creates, internet users have an ever-increasing number of web sites to choose from, including e-commerce and content sites (e.g. news sites).How to provide better ways of finding data that is interesting and useful for a user becoming the focus of researcher.

Ontologies are presented by T.R. Gruber as the backbone of semantic web applications[1]. Once the semantic knowledge is represented by ontologies, querying the semantic data and semantic searching become efficient and accurate. Semantic indexing is seeking semantic data in Resource Description Framework (RDF) knowledge by a structured way and keyword queries. A similar approach [2,3] use very basic extraction methods in RDF triples, but complex

semantics cannot be captured from the indices containing only subject predicate object triples as index terms. The Semantic Web relies heavily on formal ontologies for achieving knowledge management by capturing, sharing, and reusing knowledge within a specific domain[4]. An ontology-based approach has been proposed for a comprehensive meta-level modelling and retrieval in [5]. Creating a machine understandable ontology for a computer or software agent requires that the ontology structure be transformed into a language form such as the RDF[6].

Due to the different objectives for application, text implied sentiment analysis deals with at different levels of granularity, from document level to feature level. A text-level analysis which means the whole text expresses only one sentiment for one topic,such as[7], the derived overall polarity assessment is often not accurate. To solve this problem, a more thorough analysis at sentence level has been explored [8]. A feature-level method is presented , the method determine the semantic polarity of a Chinese word based on its morpheme and its statistical association with strongly polarized words[9].

In this paper, we presented a new approach integrated implied sentiment analysis and ontology-based method for recommendation more useful information to user. The rest of the paper is organized as follows: section 2 is on-line text pre-processing; section 3 is our recommended method; section 4 is experimental results; finally, we will conclusion our work and discuss some future work.

II. ON-LINE TEXT PRE-PROCESSING AND EVALUATION OF IMPLIED SENTIMENT METHOD

It must take some pre-processing for Chinese on-line text, such as parsing, word segmentation and part of speech tagging, because the linguistic characteristic between Chinese and English are very different. HIT LTP [10] is used to do above tasks. An open-source Chinese dependency grammar parser, called Harbin Institute of Technology Information Retrieval Lab Language Technology Platform (HIT-IR LTP) is adopted to convert each sentence into a dependency tree.

Currently, the vector space model(VSM) is the most used and the best text representation methods. The vector space model using specific weights to represent each text word

frequency. In the text vector, the value of each axis represent the importance of each term in the text[11].

For modelling extracted sentiment feature as a vector, the polarity evaluation of the sentiment feature is necessary. The polarity of a sentiment feature word mainly has two forms: prior polarity and modified polarity. Prior polarity is the general polarity of a sentiment feature word. Modified polarity is obtained from the modifier words to change intensity or reversed the polarity of the sentiment feature words. So, the method of modified polarity evaluation gets more accurate.

$$score_{c_i} = P_{c_i} - N_{c_i} \quad (1)$$

$$P_{c_i} = \frac{fp_{c_i}}{(fp_{c_i} + fn_{c_i})} * \frac{N_p}{(N_p + N_n)} \quad (2)$$

$$N_{c_i} = \frac{fn_{c_i}}{(fp_{c_i} + fn_{c_i})} * \frac{N_n}{(N_p + N_n)} \quad (3)$$

where fp_{c_i} is the frequency c_i appeared in the set of positive sentiment words; fn_{c_i} is the frequency c_i appeared in the set of negative sentiment words; N_p is the number of positive sentiment words in sentiment dictionary; N_n is the number of negative sentiment words in sentiment dictionary. As a result, if $score_{c_i} > 0$, c_i is positive; if $score_{c_i} < 0$, c_i is negative; if $score_{c_i} = 0$, c_i is neutral.

III. RECOMMENDED METHOD

A. Ontology Design

Ontology, the most popular definition is a specification of a conceptualization. Ontologies have been treated as a new form of knowledge representation for providing fundamentals of Semantic Web. The ultimate goal of the Semantic Web is driving the evolution of the current Web and automatic data discovery by enabling users to find, share, and combine information more easily.

There are different kinds of ontologies, according to the degree of formality and expressivity of the language used to describe them. The most important aspect of ontology used for Semantic Web applications is to identify two ontology layers: the ontology definition layer and the ontology instance layer.

The ontology definition layer represents real world entities, which is a framework for establishing a structure of ontology and for describing each concept(node) in it. There are two typical relations for constructing the trees which provide the backbone to these ontologies. They are is-a and part-of relations. It is worth noticing that, the is-a relation is more precisely than part-of relation, when we translated into subsumption ($\subseteq$). Therefore is-a constitutes the basic backbone of the subsumption based hierarchical structure of a

domain. Additionally, the ontology definition contains detailed descriptions of all concepts (nodes) of ontology. This is accomplished by assigning a set of different properties to concepts (nodes).

Ontology individuals can be built when the ontology definition is constructed. It means that the properties of the concepts(nodes) are initialized, datatype properties are filled out with specific values, and object properties are linked to individuals that are instances of other concepts.

Once we built up a big network of ontology, how to identify the relationship($\supset, \subseteq$ and $\perp$) between concepts(nodes) efficiently becomes a challenge. In this paper, we adopted the algorithm of minimal mapping which presented by Vincenzo Maltese and Fausto Giunchiglia on 4th ontology matching workshop 2009[12]. They proposed the use of MinSMATCH for the computation of the minimal mapping and showed that the algorithm is significantly faster w.r.t. state of the art semantic matching tool and reduces the problem of lack of background knowledge.

B. Similarity Algorithm based on Ontology and Sentiment Inclination

First, we given a description of proposed method for determining similarity between two concepts when all features of the concepts are extracted and corresponding sentiment score of concepts are calculated. Then, we used ontology minimal mapping model to obtain the relation between c_i and c_j , included($\supset, \subseteq$ and $\perp$). In such a case, the proposed similarity measure between two concepts c_i and c_j as followings:

1) When the obtained relation is $c_i \supset c_j$, then it means c_j is a subclass of c_i , the similarity formula consist of two components.

First component, $sim_1(c_i, c_j)$, represents similarity based on the features and sentiment score that are shared between two concepts;

Second component, $sim_2(c_i, c_j)$, is used to determine contributions to the overall similarity from features and sentiment score that are different for both concepts.

For assessment of similarity, we need to define some quantities. Let $d(i)$ represent a set of concepts that concept c_i is connected to in a given ontology. Also, $d(ij)$ is the set of common concepts that both c_i and c_j are connected to.

The first component sim_1 represents a contribution to the similarity between c_i and c_j determined based on direct

connections between c_i and c_j , as well as common features shared between both concepts. Because of c_j is a subclass of c_i , some shared concepts c_k can be connected to c_i and c_j with more than one relation. In such a case, $R(c_i, c_k)$ and $R(c_j, c_k)$ represents sets of relations, and we look for the maximum similarity between relations connecting concepts c_i and c_j to the shared concept c_k . Thus, $sim_1(c_i, c_j)$ is as following :

$$sim_1(c_i, c_j) = |R(c_i, c_j)| + \sum_{c_k \in d(ij)} [Rsim(r_i, r_j) \times sentis(c_i, c_j)] \quad (4)$$

where $|R(c_i, c_j)|$ is a number of direct connections between two concepts c_i and c_j ; while $Rsim(r_i, r_j)$ represents a function that evaluates similarity between relations. To better clarify Equation, if a concept c_k is connected to c_i and c_j with the same relation then the contribution due to the concept c_k is equal to one; while $sensis(c_i, c_j)$ represent the implied sentiment similarity which calculated by sentiment formula which present in section B.

$$sensis(r_i, r_j) = \begin{cases} score_{c_i} + score_{c_j} & \text{if } |score_{c_i} + score_{c_j}| < score_{c_i} + score_{c_j} \\ 1 & \text{else} \end{cases} \quad (5)$$

We can draw out from the above formula, if there is a different implied sentiment between c_i and c_j , then

sim_1 will be weakened, else sim_1 is unchanged.

The second component $sim_2(c_i, c_j)$ measures contributions to the similarity emerging from features that are unique to each concept c_i and c_j . It should be noted that

$$d_i^0 = d(i) - d(ij) \text{ and } d_j^0 = d(j) - d(ij)$$

represent unique features of each concept c_i and c_j , respectively.

The similarity is determined in the following way: for each unique feature c_z of the concept c_i with the relation

$R(c_i, c_z)$ the most similar relation between c_j and c_y from

d_j^0 is found. Next, the similarity between c_z and c_w , where

c_z is the concept connected to c_z with the most similar

relation to $R(c_i, c_z)$ is determined. This similarity

between c_z and c_w is calculated in a recursive fashion calling

$sim(c_z, c_w)$. The recursion is stopped when a pre-defined number of recursions is reached. Finally, both similarities are aggregated. The formula for $sim_2(c_i, c_j)$ is:

$$sim_2(c_i, c_j) = \sum_{c_z \in d(i)} [\max_{\substack{c_y \in d^0(j) \\ \exists R(c_i, c_z) \\ r_j \in R(c_j, c_k)}} (Rsim(r_i, r_j)) \times sentis(c_i, c_j) \oplus \max_{c_w \in d^0(j)} \{sim(c_z, c_w)\}] \quad (6)$$

Note, the component $\max \{sim(c_z, c_w)\}$ determines the highest similarity between c_z and any concept in c_w .

Finally, the similarity between concepts c_i and c_j is defined as:

$$sim(c_i, c_j) = \frac{sim_1(c_i, c_j) + sim_2(c_i, c_j)}{|d(i)|} \quad (7)$$

2) When the obtained relation is $c_i \subseteq c_j$, then c_i is a subclass of c_j , it means we found the superclass of c_i .

According to relation between superclass and subclass, c_i has more features than c_j . When c_i is a core of similarity, here we just calculated sim_1 as overall similarity. The formula as followings:

$$sim(c_i, c_j) = \frac{sim_1(c_i, c_j)}{|d(i)| * 2^n} \quad (8)$$

While n is the path distance from c_j to c_i in ontology tree.

3) When the obtained relation is $c_i \perp c_j$, it means c_i and c_j do not belong to the same ontology tree, then we will not calculated the similarity between c_i and c_j . Through execution of minimal mapping algorithm for ontology in advance, it will reduce the unnecessary calculation greatly for our presented method.

IV. EXPERIMENTAL RESULT

We compare our scenario to existing methods in determining similarity between symbolic concepts. We selected structure and semantic similarity method in order to make the comparison fair and meaningful. Rada et al. method [13]: This method counts the number of nodes along the shortest path between two objects in a structure containing the objects. Wu and Palmer method [14]: it is a structure-based measure using depth of a concept c in the structure under the study. Tversky index [15]: a degree of similarity is determined based on the features of concepts.

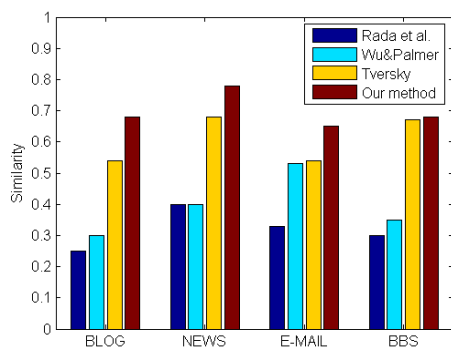


Figure 1. Comparison of similarity values with initial text for each method

These three algorithms and our proposed algorithm are applied to different on-line text field, included blog、news、email and bbs. From the figure 1. can concluded that our method always have the highest similarity values with initial text because our method not only considered ontology concepts, but also associated with implied sentiment.

Two models of recommender system are commonly used: content-based filtering(CBF) and collaborative filtering (CF).We used pure CBF、pure CF and our proposed method to filter results which search 100 questions from google and baidu search engine, then recommended results to customer. The following table shows that our proposed method has a higher clicks ration than others.

TABLE 1. THE NUMBER OF RECOMMENDED RESULTS WHICH ARE CLICKED BY CUSTOMER

Recommen dation method	No. of recommended results	No. Of Clicks	The Clicks ration
Pure CBF	958	475	49.6%
Pure CF	1100	465	42.3%
Proposed method	870	529	60.8%

V. CONCLUSIONS

This paper developed a new based-on ontology and implied sentiment similarity method to implements a Chinese on-line document recommendation. Because both implied sentiment factors and ontology structure are considered in our algorithm, the method gets a comprehensive assessment of similarity. Especially, we adopted different similarity formula according to different relation between concepts, it can reduce the execution time and improve the accuracy of algorithm.

The future research work will focus on expansion of exist ontology collection for Chinese and improvement the adaptability of algorithm.

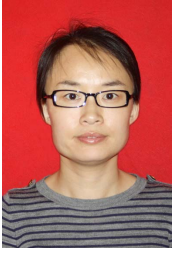
ACKNOWLEDGMENT

This paper is supported by (1) the National Natural Science Foundation of China under Grant No.608731466097309260-

903 097, (2) National High Technology Research and Development Program of China under Grant No.2007AAO4 Z114, (3) project of science and technology innovation platform of computing and software science (985 engineering), (4) the Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China. The authors would like to thank the reviewers for their valuable comments.

REFERENCES

- [1] T.R. Gruber, Toward principles for the design of ontologies used for knowledge sharing, *International Journal of Human Computer Studies* 43 (1995) 907–928.
- [2] U. Shah, T. Finin, A. Joshi, R.S. Cost, J. Matfield, Information retrieval on the semantic web, in: *CIKM '02: Proceedings of the Eleventh International Conference on Information and Knowledge Management*, ACM, New York, NY, USA, 2002, pp. 461–468.
- [3] I. Celino, E.D. Valle, D. Cerizza, A. Turati, Squiggle: an experience in model-driven development of real-world semantic search engines, in: L. Baresi, P. Fraternali, G.-J. Houben (Eds.), *ICWE, Lecture Notes in Computer Science*, vol. 4607, Springer, 2007, pp. 485–490.
- [4] Beydoun, G., Low, G., Tran, N., & Bogg, P. (2011). Development of a peer-to-peer information sharing system using ontologies. *Expert Systems Application*, 38(8),93529364. <http://dx.doi.org/10.1016/j.eswa.2011.01.104>.
- [5] M. Liao, A. Abecker, A. Bernardi, K. Hinkelmann, M. Sintek, Ontologies for knowledge retrieval in organizational memories, in: *Proceedings of the Workshop on Learning Software Organizations*, Fraunhofer Institute for Experimental Software Engineering, 1999, pp. 11–25.
- [6] Brickley, D., & Guha, R. V. (2004). RDF vocabulary description language 1.0: RDF Schema Retrieved Jan 10, 2006, from <http://www.w3.org/TR/rdf-schema>
- [7] Turney, P. D., & Littman, M.L. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4), 315-346, 2003
- [8] Meena, A., Prabhakar, T.V. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *Proceedings of the European Conference on IR Research. Lecture Notes in Computer Science*, 4425, 573-580, 2007
- [9] Yuen, R.W.M., Chan, T.Y.W., Lai, T.B.Y., Kwong, O.Y., T'sou, B.K.Y. Morpheme-based derivation of bipolar sentiment orientation of Chinese words. In *Proceedings of the 20th International Conference on Computational Linguistics*, 1008, East Stroudsburg, 2004
- [10] Ye, Q., Shi, W., Li, Y. Sentiment Classification for Movies Reviews in Chinese by Improved Semantic Oriented Approach. In *HICSS'06: Proceedings of the 39th Annual Hawaii International Conference on System Sciences* 53.2 2006.
- [11] Hui Yang, Zhe Wang, Xu Zhou , Zuo Wang, Tong Zhou. A Chinese online document clustering algorithm based on hidden sentiment vector, *Journal of Computational Information Systems*, 2011, 1974-1987
- [12] V. Maltese, F. Giunchiglia, A. Autayeu, "Save Up To 99% of Your Time in Mapping Validation", 4th Ontology Matching Workshop, 2009
- [13] R. Rada, H. Mili, E. Bicknell, Development and application of a metricon semantic nets, *IEEE Transaction on Systems, Man and Cybernets* 19 (1989) 17 – 30
- [14] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, Las Cruces, New Mexico, 1994, pp. 133 – 138.
- [15] A. Tversky, Features of similarity, *Psychological Review* 84 (7) (1977) 327 – 352.



Yang Hui obtained her Ph.D. in computer science from JILIN University in 2011. Her research interest is social network Mining ,sentiment analysis and information retrieval. She is currently an lecturer at the department of Computer Science in Jilin University .She researched Multi-language UKC as a post-doc in the university of Trento(Italy).



Wang Zhe is corresponding author. He is an associate professor, and master's supervisors in Jilin University. Mainly engaged in the research of business intelligent, data mining, and social network, taking in charge of one National science and technology support plan corpus, one key project of Jilin Province Science and Technology Agency, and one general project.



Hao Xu is an associate professor at college of computer science and technology, Jilin University, China and a courtesy professor of KnowDive research group, University of Trento, Italy. He received his PhD in Information and Communication Technologies (ICT) in University of Trento, Italy. His research interests include knowledge management, data analytics and human engagement.