Semantic Annotation for Dynamic Web Environment

Jeong-Hoon Park
Department of Computer Science, KAIST
291 Daehak-ro, Yuseong-gu
Daejeon, Korea
jhpark@islab.kaist.ac.kr

Chin-Wan Chung
Division of Web Science and Technology &
Department of Computer Science, KAIST
291 Daehak-ro, Yuseong-gu
Daejeon, Korea
chungcw@kaist.edu

ABSTRACT

The semantic Web is a promising future Web environment. In order to realize the semantic Web, the semantic annotation should be widely available. The studies for generating the semantic annotation do not provide a solution to the 'document evolution' requirement which is to maintain the consistency between semantic annotations and Web pages. In this paper, we propose an efficient solution to the requirement, that is to separately generate the long-term annotation and the short-term annotation. The experimental results show that our approach outperforms an existing approach which is the most efficient among the automatic approaches based on static Web pages.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Semantic Annotation, Dynamic Web

1. INTRODUCTION

As the amount of information in the Web has been increasing continuously, intelligent methods for effectively manipulating the information in the Web are mandatory. The semantic Web is an intelligent Web where data about Web resources is formally represented.

In spite of the advance of the Web, the simple hyper-link is still used and Web contents are described in the plain text format. In order to utilize the Web contents in the semantic Web, the semantic annotation is proposed. The semantic annotation is a set of semantic labels assigned to a Web page, which are ontology concepts(classes and instances) corresponding to the information the Web page presents.

These days, the contents of Web pages are dynamically updated and configured in order to intelligently present the up to date information. Therefore, the 'document evolution' requirement introduced in [3] is one of the most important requirements for the semantic annotation. The requirement is to maintain the consistency between Web pages and semantic annotations. There have been many studies[2, 1] related to the semantic annotation. However, the studies are not focused on the 'document evolution' requirement.

In this paper, we propose a solution to the requirement. The solution is separating the semantic annotation into the long-term annotation and the short-term annotation. The long-term annotation is the semantic annotation which is consistent with a Web page for a long time by describing static information of the Web page. The short-term annotation is the semantic annotation which is prone to inconsistency by describing the dynamic information of the Web page. Even though some parts of a Web page are dynamically updated, the consistency of the long-term annotation is rarely affected by the updates. In contrast, the consistency of the short-term annotation is affected by the updates. As a result, the system can focus on updating the short-term annotation in order to efficiently maintain the consistency. In order to separately generate the long-term annotation and the short-term annotation, we use anchor texts and contents of Web pages, respectively. An anchor text is the visible and clickable words in a hyperlink which provides the summary description of the target Web page. The experimental results show that our approach outperforms an existing approach which is the most efficient among the automatic approaches based on static Web pages.

2. LONG-TERM ANNOTATION

In this section, we present a method to generate long-term annotations of Web pages. We consider anchor texts as the source of the long-term annotation in order to efficiently extract the static information from Web pages. An anchor text for a Web page u_d is a set of keywords that are clickable in another Web page and linked to u_d. An anchor document of u_d is the union of all the anchor texts for u_d. Therefore, the goal is generating long-term annotations from anchor texts of Web pages. Our method for generating long-term annotations consists of 3 steps as follows:

[$Step1$:Elimination of Noise Words] Among the keywords in the anchor text, this step selects only the top m keywords for eliminating noise words. For the top m selection, given a keyword k_j in a document u_d, we propose the $TFIAF$ measure for scoring each keyword as follows:

$$TFIAF(k_j, u_d) = TF(k_j, u_d) \cdot IAF(k_j)$$

where $TF(k_j, u_d) = \frac{|\{a_i | a_i \in AT_d, k_j \in a_i\}|}{|AT_d|}$, and $IAF(k_j) = \log\left(\frac{|AD|}{|\{AD_d | k_j \in AD_d, AD_d \in AD\}|}\right)$, AT_d is the set of anchor texts for u_d, and AD is the set of anchor documents for all $u_d \in U$.

[$Step2$:Extracting Candidate Semantic Concepts] In this step, we extract candidate ontology concepts which are likely to have the semantic relevance with the selected words. For the extraction, we utilize the $Jaro\text{-}Winkler(JW)$ edit distance [4] which gives more weight to the texts sharing the common prefix. Among the semantic concepts and a given selected keyword, we choose the semantic concepts such that, for each concept, the JW edit dis-

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577373.

tance similarity between the name and the given selected keyword is greater than γ. We find that 0.78 is the optimal values for γ by our empirical study. We denote the set of candidate semantic concepts of k_t as $CAND_t$.

[$Step3$:Long-term Word Sense Disambiguation(LWSD)] In this step, we select the most relevant concept among the candidate semantic concepts for each selected keyword. For given k_t in the selected anchor keywords of u_d, and a concept $c_{t,p} \in CAND_t$, the relevance between k_t and $c_{t,p}$ is computed as

$$L^d_{t,p} = \sum_{d_c \in CD(k_t,u_d)} WN\left(d_c, C^c_{t,p}\right) \cdot SR\left(C^c_{t,p}, c_{t,p}\right)$$

where $CD(k_t, u_d)$ is the set of u_d's anchor texts which contain k_t, $WN\left(d_c, C^c_{t,p}\right)$ is the value of weighted naive bayesian classifying of d_c w.r.t. $C^c_{t,p}$, $C^c_{t,p} = \arg_{c_i \in SC(c_{t,p})} \max(WN(d_c, c_i))$, $SC(c_{t,p})$ is the semantic context of $c_{t,p}$ which consists of the neighbor concepts of $c_{t,p}$ in the ontology, $SR(c_1, c_2)$ is the semantic relevance between c_1 and c_2 which is computed by $SR(c_x, c_y) = \frac{2 \cdot len(r,c_p)}{len(c_x,c_p)+len(c_y,c_p)+2 \cdot len(r,c_p)}$ where r is the root concept, c_p is the least common ancestor concept and len is the distance between two input concepts.

3. SHORT-TERM ANNOTATION

In this section, we present a method for generating short-term annotations based on the long-term annotations generated in advance. Since the long-term annotation of a Web page corresponds to the static information of the Web page, the short-term annotation should be semantically related to the long-term annotation.

The short-term annotation of a Web page u_d is based on the texts contained in u_d. The method for generating the short-term annotation contains the three steps as the method for the long-term annotation. Instead of $TFIAF$, the $TF\text{-}IDF$ measure is used in $Step1$, and the same method based on the JW edit distance is used in $Step2$. But the input texts are from the contents of u_d. In $Step3$, in order to choose top one concept for each selected keyword, we devise a relevance model for scoring the concepts. Given a Web document u_d, the u_d's long-term annotation LT_{u_d}, and a candidate semantic concept $c_{t,p}$, the relevance model is

$$AR(c_{t,p}, LT_d) = \sum_{c_i \in LT_d} CNT(c_i, LT_d) \cdot SR(c_i, c_{t,p})$$

where $CNT(c_i, LT_d) = \sum_{c_j \in LT_d, c_j \neq c_i} \frac{SR(c_i, c_j)}{|LT_d| - 1}$.

$CNT(c_i, LT_d)$ denotes the closeness centrality of c_i in LT_d. The more central a concept is, the more important the concept is in LT_d.

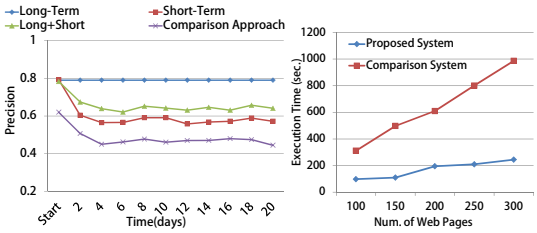
4. EXPERIMENTS

In order to validate our approach, we implement the semantic annotation system based on our approach and conduct experiments using a real data set(real Web pages, YAGO ontology and DBpedia).

In order to use real data in the experiments, we crawl about 1.8 millions of real Web pages in famous sites such as imdb.com, weather.com, twitter.com, and amazon.com. The number of link relationships is about 9 millions. The YAGO ontology contains about 2 millions of concepts and about 21 millions of triples. In the Yago ontology, concepts and triples are constructed from Wikipedia and WordNet. We focus on validating the precision and the generation time of our approach. Our approach is compared with [1] that is based on 'Part Of Speech'(POS) and the text similarity between words in Web pages and names of the ontology concepts. [1] is the most efficient among the automatic semantic annotation systems targeting static Web pages.

4.1 Precision

We compare the precisions of the long-term annotation and the short-term annotation generated by our approach, and the semantic annotation generated by the comparison approach. The precision of a Web page is measured by the ratio of the true positives among the concepts in the semantic annotation of the page.



(a) The Precision over Time (b) The Generation Time

Figure 1: The Experimental Results

Fig. 1(a) shows the precisions of our approach and the comparison approach for 20 days. The precisions of our approach are better than those of the comparison approach. The word sense disambiguation of the comparison approach is based on the text matching algorithm, and the comparison approach does not eliminate noise words. These points yield the lower accuracy. In addition, As we can see, the long-term annotation is stable for 20 days. This shows that the long-term annotation generated by our approach is accurate even though the contents of the Web pages are frequently updated.

4.2 Efficiency

The semantic annotations for a huge amount of Web pages should be updated in order to maintain the consistency. Therefore, the generation time affects the maintenance of the quality of the semantic annotation to be used by various applications.

Fig. 1(b) shows the generation times of our approach and the comparison approach according to the number of Web pages. The average number of annotations for each Web page is about 320. The generation time of our approach is about 4 times faster than that of the comparison approach.

5. CONCLUSION AND FUTURE WORKS

In this paper, we propose a method to separately generate the long-term annotation and the short-term annotation as the solution to the 'document evolution' requirement. The method for generating the semantic annotation consists of 3 steps: elimination of noise words, extraction of candidate ontology concepts, and word sense disambiguation. The experimental results show that our approach outperforms an existing approach.

As a future work, we expect that Step2 can be improved by adapting advanced IR techniques. Step2 affects the efficiency and accuracy of the proposed method since Step2 prunes irrelevant semantic concepts.

Acknowledgements. This work was supported by the National Research Foundation of Korea grant funded by the Korean government (MSIP) (No. NRF-2009-0081365).

6. REFERENCES

[1] Köhler,, Jacob and Philippi,, Stephan and Specht,, Michael and Rüegg,, Alexander. Ontology based text indexing and querying for the semantic web. *Knowledge Based System*, 19(8):744–754, December 2006.

[2] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM - Semantic Annotation Platform. In *Proc. of 2nd International Semantic Web Conference (ISWC2003)*, pages 834–849, 2003.

[3] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14–28, January 2006.

[4] W. E. Winkler. The state of record linkage and current research problems. In *Technical report, Statistical Research Division, U.S. Census Bureau, Washington, DC*, 1999.