The International Journal Of Engineering And Science (IJES)
|| Volume || 6 || Issue || 1 || Pages || PP 12-16 || 2017 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Application of Semantic Tagging to Academic Paper Services

Sumi Shin
S&T Information Center, Korea Institute of Science and Technology Information

ABSTRACT

As the semantic web becomes more common, interactive data have gotten more important. In academic papers which provide online publish services, therefore, their usage and strength can be enhanced by applying semantic tagging. Semantic tagging makes it possible to understand the characteristics and categories of tagged words at a glance and use them for visualization. If detailed explanation on the tagged words can also be viewed at the same time while reading a paper, in addition, readers' convenience and legibility can be improved simultaneously. This study compared the keywords tagged in Spotlight and TAGME respectively and those tagged in both systems. Among 10 papers in 5 subjects, those with improved keyword matching accounted for 70%, 60%, 50%, 60% and 80% respectively. In all topics, Spotlight was higher than TAGME in terms of average index, revealing a similar pattern. Therefore, it was estimated that there would be a correlation between Spotlight and TAGME by topic. Then, improvement in performances when Spotlight and TAGME are integrated would be examined. If tagged including TAGME instead of using Spotlight only, improved performances are expected. Using more appropriate resources for each domain as LOD, furthermore, it should be able to tag more diverse words in more accurate fashion in addition to DBpedia Spotlight and TAGME. Then, the efficiency of development and tag results could be enhanced by sorting and clearly classifying diverse resources simultaneously.

Keywords: *Academic paper services, Semantic Tagging, Semantic Web, Tagging*

Date of Submission: 22 December 2016 Date of Accepted: 05 January 2017

I. INTRODUCTION

The next generation web services are recognized as the next generation of web technologies in web 3.0-based services from the semantic web[1]. One of the most critical elements in recent semantic information services is tagging. So far, there have been a lot of studies on tagging from diverse aspects. The famous websites commonly used in studying tagging are Yahoo's photo storage site 'Flickr (http://flickr.com/)' and the social bookmarking site 'Delicious (http://del.icio.us/).' 'YouTube (http://youtube.com/)', the world's popular video-sharing website, is also one of the systems in which tagging is useful. Since any technology which can read the context in a video is not available yet, users estimate the contents by the tags they attach. Before tagging was introduced, videos were just searched by title. However, it is not enough to describe a video which contains a lot of information compared to texts or pictures with a short title only.

Semantic tagging refers to tagging for semantic web. The Semantic Web is a collaborative movement led by the international standards body, the World Wide Web Consortium (W3C). The purpose of this web technology is to make data sharing, search and fusion easier than current web. The Semantic Web is a system specially designed to understand and properly respond to the diverse and complicated human requests. Shotton defined semantic publishing as a series of efforts to enable the mutual semantic connection between journals, to enable approaches to data within a journal, or to enrich the meaning of published journals[2]. Also, according to Shotton, semantic publishing was defined as a series of activities that connect semantically related papers or enable an approach to data within a paper. In Wikipedia, semantic publishing means to provide a method of understanding the structure and meaning of information published by computer (or agent) by publishing information on the website in a document form including semantic markup, and to make published information searched and integrated more efficiently.

The conventional World Wide Web usually uses HTML texts so that it's been unable to find the exacting for the context. In addition, they have been hardly recognized as objects which differ from other words in the page. Ontology engineering in Semantic Web is primarily supported by languages such as RDF, RDFS and OWL[3]. Therefore, the Semantic Web adopts a language specially designed for data such as Resource Description Framework (RDF), Web Ontology Language (OWL) and Extensible Markup Language (XML). In other words, semantic tagging is tagging on external resources and particular words, using the RDF or OWL. Web is ontology that is used to explicitly represent our conceptualizations. Ontologies are built to model a domain and support reasoning over the concepts[4].

II. RELATED WORKS

One of the controversies associated with tagging is how much search efficiency could be enhanced immediately. From a positive point of view, several tags can be placed on one object because multiple tags which represent a single object in diverse fashion broaden a scope of search. In fact, this kind of contention has been suggested as the result of studies on library science since 1980s. At that time, however, current web environment didn't exist. Therefore, a concept of 'social tagging' in which several users participate wasn't available either. This tag-adding process constitutes the frame of a folksonomy, partially playing the role of collective intelligence.

In tagging studies, linguistics and computational linguistics account for a great portion as well. What is interesting in computational linguistics-related tagging tasks is a subject tag so-called 'meta keyword'[5]. In a meta keyword system, users are able to watch a video and express how they felt such as "fun" and "boring" using tags. These meta keywords bring emotion and vigor to the contents, but they could rather drop search efficiency due to too much complicated tag spaces. Therefore, there have been studies on structuring tag spaces from the other aspect. The techniques which create the same effects through the studies which derive meaning from Flickr tags and tag clustering also originated from a similar motive. The ontology is constructed using the inference technology of artificial intelligence. In tag clustering is adopted clustering algorithm which has become base technology in natural language processing.

As the realtime processing of large data is enabled due to the development of artificial intelligence technology, there have been a lot of linguistic studies such as statistical analysis, survey on tag consumption and distribution of tag vocabularies. Under this kind of trend, statistical techniques are introduced to computational linguistics, bringing an innovative turning point in several fields such as machine translation. The studies in which tagging is understood as a process of constructing collective knowledge which is expressed in particular language belong to this category. With the introduction of social tagging, tagging has occurred simultaneously because of several users. As a result, user behavior has become important in studies on tagging as well because tagging is now a collaborative and dynamic phenomenon just like the automated recommendation system, not a static, one-time operation. In Germany, recently, there has been a study on the science paper bookmarking system 'Connotea,' not Flickr or Delicious designed for general users. The University of Regensburg analyzed users' tagging type against Connotea which secured the largest number of users in science paper along with CiteULike.

The study empirically analyzed if users' tagging types can be normalized and what are the differences between author-selected keywords and social tagging. One of the interesting aspects is that it investigated if the Internet short-hands or cyber-slangs which are commonly used in an online community occur in a tagging community as well. The cyber slangs evolve in diverse fields such as chat room, email text, messenger chat window, text message and blog, and a large portion would be overlapped. Sometimes, they would happen in a certain community just like DC Inside and develop into the Internet culture. In Connotea tagging which is aimed to share academic papers, highly educated users who are more conservative than the youth who frequently use new communication words in using language participate. In a tagging type, even so, various language usages (e.g., omission of a punctuation mark, no classification of small and big letters, creation of compound words, etc.). These results suggest that computational approach would be requested continuously even in tagging fields where grammar and semantics are relatively simple.

One of the most important applications regarding social tagging is tagging in corporate applications[6]. This field has a limitation of internal users only, not unlimited general users. It earned high scores in practicality as well. The most active staff tagging technology in corporation broadens the scope of utilizing human resources by attaching human information such as technology and expertise in tagging fashion. Some studies have constructed a tagging application system for business, which combines multi-user platform and task prioritizing algorithm[7].

DBpedia Spotlight (hereinafter "the Spotlight") is a tool for detecting the mentions of DBpedia resources in the text. The Spotlight which is available as a web service or Java/Scala API for the purpose of testing links text documents with linked open data (LOD) using DBpedia as an interlinking hub. As of September 2013, the number of interlinks including the external data set with the reference DBpedia 'Freebase,' OpenCyc, UMBEL, GeoNames, Musicbrainz, CIA World Fact Book, DBLP, Project Gutenberg, DBtune Jamendo, Eurostat, Uniprot, Bio2RDF and US Census data exceeds 45 million.

Spotlight performs spotting using the labels of the extended DBpedia lexicalization datasets. In terms of Wikipedia dataset, a named entity is disambiguated, using vector space model (VSM). Unlike the conventional TF-IDF weighting scheme, Spotlight introduced TF-Inverse Candidate Frequency (ICF) weighting and revealed that TF-ICF weighting is better than TF-IDF weighting in terms of performance. A dictionary which includes the surface form of Freebase Id is constructed, and unnecessary types were removed from a list of stop-words after reading the given texts and finding mentions using the longest match strategy. After connecting the most related entities in each track (long track and short track), best candidates were extracted. Then, great performance was achieved through experiences by combining Spotlight with TAGME.

In a short track, the system can be improved by returning N-best candidates or all other candidates which exceed the predefined threshold. In a long track, in contrast, the system needs to be improved through disambiguation, using the contextual information which is found in Spotlight and TAGME from the short track.

TAGME is a technology which hyperlinks plain-texts with appropriate Wikipedia pages through the system proposed by Ferragina and Scaiella in 2010. TAGME is special in that it is able to annotate in short writings with poor construction ability such as snippet of search results, tweet and news.

In basic idea, the studies written by Milne and Witten were followed. Instead of considering the relatedness of unambiguous keywords only, a method which analyzes the relatedness of the trained ambiguous keywords was introduced. In fact, it was able to achieve better results and maintain the strength of long texts when a short text was focused under this method.

III. PROPOSED WORK

An abstract is an abbreviated representation of an academic paper with short texts so that it is not easy to figure out the characteristics of each domain (subject) with this brief summary only. Therefore, this paper was able to review each subject's features with keywords and indexes. If the words tagged by applying semantic tagging to the entire paper are categorized, clear features by domain would appear. In terms of the number of average keywords by journal, the fields other than 'medicine' include 6.5 thru 9.2 words. However, 'medicine' has 17.7 keywords in average. It was forecasted that the number of keywords would increase as a text becomes longer because of the length of abstract. In addition, the size of abstract texts was compared. In 'agriculture' journals, fungi are mostly tagged. Considering the fact that the length of these words is longer than that of the words from other fields, text size and number of keywords are proportional to each other. This hypothesis could be clearly proven by comparing the length of keywords by academic paper.

This study compared the annotation tools 'Spotlight' and 'TAGME' in order to apply semantic tagging to academic papers. The two systems are able to control the level of tagging with confidence and rho each. The two values range from '0' to '1.' In this experiment, they were measured at every 0.05 intervals from '0' to '1.'

This study intended to examine confidence and rho values which are good for the characteristics of two systems and each paper's subject before they are applied to the entire paper. The purpose of this test is to find out index relations and characteristics of subject between each subject and the two systems after designating an appropriate value as an index before they are applied to the paper.

Among academic papers, five subjects were chosen. After selecting one among each subject's journals, a total of 50 were sampled from 10 papers as shown in TABLE1.

Then, they were evaluated with the following steps:

1) Select the keywords which are appropriate for each paper's abstract and save them in the corresponding folder under the name of 'keyword.txt.'
2) Increase each paper's abstract confidence and rho by 5% from '0' to '1' using the demo version and API of Spotlight and TAGME and save the annotated results in 'output.txt.'
3) In each confidence and rho, the value of the point in which the absence of keywords didn't exceed 30% in each confidence and rho was set as the index (confidence and rho to be applied) of each paper. In other words, the maximum confidence and rho in which the tagged words are more than 70% among the keywords were chosen. In case of TAGME, however, if there is no index which exceeds 70%, the index was set to 0.0.
4) The average of the indexes in 10 academic papers was summarized with average indexes.

In five categories, Spotlight ranged from 0.42 to 0.72 as shown in TABLE2.

Table 1. Sample Articles

No.	Title
1	Antioxidant and Antibacterial Activities of Chitosan-Phloroglucinol Conjugate
2	Antimicrobial Resistance and Virulence Genes Presence in Escherichia coli Strains Isolated from Gomso Bay, Korea
3	New Occurrences of Two Penaeid Species (Crustacea: Decapoda: Dendrobranchiata) in Korean Waters
4	New record of Juvenile Stethojulis trilineata (Perciformes: Labridae) from Korea, Revealed by Molecular Analysis
5	Molecular Cloning, Purification, and Characterization of a Cold-Adapted Esterase from Photobacterium sp. MA1-3
6	Development of Polymorphic Microsatellite Markers Suitable for Genetic Linkage Mapping of Olive Flounder Paralichthys olivaceus
7	Genomic Organization, Tissue Distribution and Developmental Expression of Glyceraldehyde 3-Phosphate Dehydrogenase Isoforms in Mud LoachMisgurnus mizolepis
8	Effect of High Stocking Rates on Growth and Survival of the Endangered Rio Grande Silvery Minnow Hybognathus amarus
9	Distillers Dried Grain from Makgeolli By-product Is Useful as a Dietary Ingredient for Growth of Juvenile Sea Cucumber Apostichopus japonicus
10	Expression of c-Type Lysozyme from the Fleshy Shrimp Fenneropenaeus chinensis Is Upregulated Following Vibrio anguillarum and Lipopolysaccharide Injection

Table 2. Spotlight's Average index of five Subjects

Subject	Journal	Average
Agriculture	Fisheries and Aquatic Sciences	0.73
Bio and Life Science	International Journal of Industrial Entomology	0.525
Engineering	International Journal of Aeronautical and Space Sciences	0.47
Earth, Ocean and Environment	Environmental Engineering Research	0.42
Medicine	Journal of Pharmacopuncture	0.485

While TAGME was 0.14 thru 0.37 in terms of average index range as shown in TABLE 3.

Table 3. TAGME's Average index of five Subjects

Subject	Journal	Average
Agriculture	Fisheries and Aquatic Sciences	0.37
Bio and Life Science	International Journal of Industrial Entomology	0.185
Engineering	International Journal of Aeronautical and Space Sciences	0.14
Earth, Ocean and Environment	Environmental Engineering Research	0.145
Medicine	Journal of Pharmacopuncture	0.175

Even though confidence and rho have the value from '0' to '1,' confidence is always higher than rho in terms of the index which reveals 70% of keyword set. The index comparison by domain and article between Spotlight and TAGME is more specifically stated in the items. The portion of nouns among tagged words was examined when TAGME and Spotlight were applied for an optimum index. The portion of nouns between the two systems ranged from 69.6 to 91.5% in average as shown in Table4.

Table 4. Ratio of nouns in tagged words

	Agriculture		**Bio and Life Science**		**Engineering**		**Earth, Ocean, and Environment**		**Medicine**	
	Tagme	Spotlight	Tagme	Spotlight	Tagme	Spotlight	Tagme	Spotlight	Tagme	Spotlight
1	0.625	1.000	0.833	1.000	0.786	0.778	0.533	0.667	0.917	0.957
2	0.700	0.917	1.000	0.923	0.625	0.625	1.000	0.929	0.739	0.765
3	0.750	0.700	0.875	0.800	0.850	0.900	0.762	0.762	0.826	0.773
4	1.000	1.000	0.688	0.741	0.763	0.846	0.690	0.667	0.700	0.857
5	0.947	0.923	1.000	0.818	0.786	0.833	0.708	0.692	0.882	0.889
6	0.714	0.833	0.781	0.727	0.857	1.000	0.923	0.900	0.917	0.917
7	0.818	1.000	0.846	0.929	1.000	0.900	0.706	0.731	0.654	0.860
8	0.000	1.000	0.800	0.667	0.767	0.857	1.000	1.000	0.729	0.889
9	0.714	0.875	0.889	1.000	1.000	1.000	0.750	0.857	0.769	0.714
10	0.686	0.900	0.867	1.000	0.702	0.833	0.720	0.773	0.862	0.840
Avg	0.696	0.915	0.858	0.860	0.814	0.857	0.779	0.798	0.800	0.846

Then, stop-words were excluded from the results annotated after applying the index estimated at a sampling test to each journal. Then, how much they were improved was investigated by comparing the portion of the annotated keywords when Spotlight was applied, and both Spotlight and TAGME were applied at the same time. Table 5 states the union of annotated keywords, non-overlapped ratio with the annotated keywords of Spotlight among the ones in TAGME and a portion of the annotated keywords of Spotlight in Spotlight and TAGME in engineering category.

Table 5: Keyword stats in engineering category

	Spotlight + Tagme	**Tagme only keyword**	**Spotlight only keyword**
Engineering	100.0%	16.7%	83.3%
	87.5%	0.0%	87.5%
	83.3%	0.0%	83.3%
	76.9%	0.0%	76.9%
	87.5%	12.5%	75.0%
	75.0%	0.0%	75.0%
	80.0%	0.0%	80.0%
	100.0%	16.7%	83.3%
	85.7%	14.3%	71.4%
	90.9%	9.1%	81.8%

In most fields other than 'agriculture,' there was no big difference between the two systems. Then, the portion of stop-words among the tagged words was investigated when TAGME and Spotlight were applied for an optimum index. In most fields other than ‘engineering,’ TAGME was lower than Spotlight in terms of the importance of stop-words. Even so, no big difference was found between the two systems, and average importance was 6% or less.

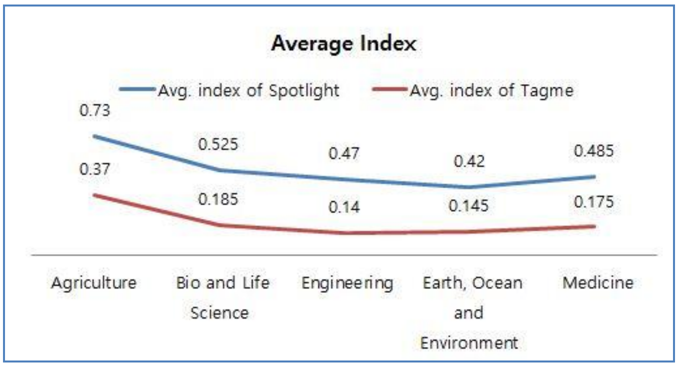

Figure 1. Average Index of Spotlight and TAGME

IV. CONCLUSION

As the semantic web becomes more common, interactive data have gotten more important. In academic papers which provide online publish services, therefore, their usage and strength can be enhanced by applying semantic tagging. Semantic tagging makes it possible to understand the characteristics and categories of tagged words at a glance and use them for visualization. If detailed explanation on the tagged words can also be viewed at the same time while reading a paper, in addition, readers’ convenience and legibility can be improved simultaneously. This study compared the keywords tagged in Spotlight and TAGME respectively and those tagged in both systems. Among 10 papers in each subject, those with improved keyword matching accounted for 70%, 60%, 50%, 60% and 80% respectively. In other words, if tagged including TAGME instead of using Spotlight only, improved performances are expected. Using more appropriate resources for each domain as LOD, furthermore, it should be able to tag more diverse words in more accurate fashion in addition to DBpedia Spotlight and TAGME. Then, the efficiency of development and tag results could be enhanced by sorting and clearly classifying diverse resources simultaneously. Furthermore, it is needed to have readers get the information they want easily by adding the function which enables the classification of tagged words by setting categories.

The examples of PLOS NTD are the results of manually tagging, not automatic tagging, and they suggest a future direction. Even though conventional automated techniques have a lot of limitations, it is needed to set a direction according to the purpose of academic paper services and realize them through diverse technologies. Since semantic tagging can improve visualization and data readability at the same time, there should be diverse attempts and technology applications. Then, stop-words were excluded from the results annotated after applying the index estimated at a sampling test to each journal. Then, how much they were improved was investigated by comparing the portion of the annotated keywords when Spotlight was applied, and both Spotlight and TAGME were applied at the same time.

REFERENCES

[1]. Ji-moon Chung, A Study of Future Internet Applications based on Semantic Web Technology Configuration Model. *Indian Journal of Science and Technology, August, 8 (19),* 2015, pp. 1-5.

[2]. Shotton. D, Semantic publishing: the coming revolution in scientific journal publishing, *Learned Publishing, 22(2),* 2009, pp. 85-94..

[3]. Ding L, Kolari P, Ding Z, Avancha S. Using ontologies in the semantic web: A survey. *In Ontologies. Springer US,* 2007, .p.79–113.

[4]. M. P. S. Bhatia , Akshi Kumar and Rohit Beniwal, Ontologies for Software Engineering: Past, Present and Future. *Indian Journal of Science and Technology. March, 9 (9), 2016.*

[5]. Halpin, Harry and Robu, Valentin and Shepherd, Hana, The Complex Dynamics of Collaborative Tagging, *Proceedings of the 16th International Conference on World Wide Web,* 2007, 221-22-.

[6]. Social Tagging For The Enterprise. http://imbok.blogspot.kr/2006/02/social-tagging-for-enterprise.html.

[7]. Ajita John , Doree D. Seligmann, Collaborative tagging and expertise in the enterprise, *PROCEEDINGS OF COLLABORATIVE WEB TAGGING WORKSHOP HELD IN CONJUNCTION WITH WWW* 2006

S.M Shin received the B.S. and M.S. degrees in Computer Engineering from Hongik University, Korea, in 1997 and 2005, respectively. She is a Ph.D. student in Dept. of Computer Engineering, Hongik University, Korea. And she is currently a senior researcher in the Department of Information Service, KISTI, Korea. She is interested in recommender system, semantic web, big data analysis, databases, data retrieval and Internet of Things.