ISSN 1749-3889 (print), 1749-3897 (online)
International Journal of Nonlinear Science
Vol.18(2014) No.1,pp.86-91

Ontology Evolution Algorithm for Topic Information Collection

Jing Ma[1] *, Mengyong Sun[1], Chi Li[2], Zihao Tian[1]
[1] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2] College of Mathematics, University of Science and Technology of China, Hefei, China

(Received 10 March 2014, accepted 08 July 2014)

Abstract: In this paper, we propose a domain ontology evolution algorithm based on web-pages and user-behavior, which can excavate new concepts dynamically and realize topic-ontology evolution under users' guidance. The new ontology will be applied to direct topic information collection, which is a virtuous circle and the whole subject information collecting system would optimize automatically and continuously. The results of experiment verify that this solution could effectively improve accuracy and coverage-rate of topic information collection.

Keywords: topic information collection; domain ontology; ontology evolution

1 Introduction

Topic information collection is means collecting web-pages that are related to a pre-specified subject. Ontology Evolution means according to theories, technology and standards, enrich and complete concepts and relationships step by step on the basis of original core ontology [1]. And the difficult point is how to extract new concepts and relationships between concepts. Bourigault [2] thinks that there is a fixed lexical form for term, and extracts noun phrases with max length based on shallow grammar parsing which detects word boundaries in sentences of POS tagged text, discovers grammar relationship between words to determine domain concepts. Sabou [3] puts forward a concepts mining solution relying on grammar. He takes comprehensively utilization of morpheme, root and words' position in sentences to extract concepts, most of which are noun phrase. (2) Domain concepts extraction based on statistics. Navigli [4] proposes to bring in domain irrelevant corpus to screen out concepts with domain features by comparison. Refs. [5] and [6] Discover correlation between words on word matrix and cluster keywords with social network analysis method to mine core keywords. (3) Combine two methods above together. Wu et al. [7] use PAT-Tree to find high frequency words of domain and applies them as seed words to pick out concepts related from POS tagged Chinese corpus under extraction rules. He et al. [8] use N-Gram to extract domain terms, and discard noise data generated by N-Gram. Then measure subject relativity strings with GF/GL to extract concepts. Taking a comprehensive view on ontology evolution related researches, we find that original domain ontology is usually provided by users, which could not work well with the rapid growth of information and lead to the leaking of information. Meanwhile Ontology evolution lacks usage of existing ontology. Precision of concepts extraction and relationships extraction is low. To address this problem, a domain ontology evolution solution is put forward in this paper. This evolution solution could excavate new concepts dynamically based on web-pages and user-behavior, and realize domain-ontology evolution under users' guidance. And evolved subject-ontology will be applied to direct subject information collection. This is an virtuous circle and the whole subject information collecting system would optimize automatically and continuously.

2 Ontology evolution algorithm based on web-pages and users' behavior logs

2.1 Calculating location weight of keyword

On a web-page, text between title-tag stands for the title of HTML document, which is usually displayed on title-bar of browser, or used as title of webpage bookmarks in favorite. Description attribute of meta-tag, which offers a brief de-

*Corresponding author. *E-mail address*: search@nuaa.edu.cn

scription about content of web-page, could express theme of current page. And keywords, another attribute of meta-tag, which contains words separated by space or comma, is used for classifying web-pages by search engine. The main content of page is placed between body-tag. As we know, different keywords on different position of web-page have different functions, which means they have different weights. Weight of words on meta-keywords, title-tag, meta-description and body-tag should be decreasing. Defining location-weight of keyword as $\omega_{location}$ according to the location of keyword in web-page, then $\omega_{location} = \alpha_{\omega}title + \beta\omega_{description} + \gamma\omega_{keywords} + (1 - \alpha - \beta - \gamma)\omega_{body}$ where four weights $\omega_{title}, \omega_{description}, \omega_{keywords}$ and ω_{body} are title location weight, meta-description location weight, meta-keywords location weight and body location weight of keyword, separately. And parameters ω, β and γ should meet conditions: $0 < \alpha < 1, 0 < \beta < 1, 0 < \gamma < 1$ and $0 < \alpha + \beta + \gamma < 1$. According to research results and practical experience, prevalues of α, β and γ are 0.30, 0.25 and 0.30. The following is computional formular of $\omega_{title}, \omega_{description}, \omega_{keywords}$ and ω_{body} :

$$\omega_l = \frac{1}{n_l}\sum_{i=1}^{n_l}(tf(d_i) \times log(\frac{N_l}{n_l})), \tag{1}$$

where l stands for the location of keywords, like: title, meta-description, meta-keywords, body. And d is web-page which has content on location l , N_l is the quantity of web-pages which have content on location l , and n_1 is the quantity of web-pages which contains keyword t on location l .

2.2 Calculating behavior weight of term

Besides the location weight, users' behavior weight should also be considered when calculate recommend-weight of keyword. Users' behavior factors contains subject web-pages browsed by user, search words on subject, etc. Weight of keywords contained in pages clicked by users or used to search by users should be increased properly. Depending on users' behavior tendency in real application, define behavior weight of keyword as:

$$\omega_{behavior} = \lambda\frac{SN + Sn}{SN \times Sn}\sum_{i=1}^{C_n}(tf(d_i) \times log(\frac{CN}{Cn})), \tag{2}$$

where SN is totality of words used to search by users, Sn is totality of keywords used to search by users, CN is totality of web-pages clicked by users, Cn is totality of web-pages that contain keywords and are clicked by users. d is web-page which is clicked by users, and , value of λ is 1 while d contains keywords, otherwise, 0. Therefore, users' behavior weight of web-pages which are clicked by users but do not contain keywords is 0.

2.3 Recommending new ontology classes

Considering the impact of location and users behavior factors on keywords, define recommend weight of keywords as following:

$$\omega_{recommend} = t\omega_{location} + (1 - t)\omega_{behavior}, \tag{3}$$

where t is an adjustable parameter and meets the condition that $0 < t < 1$. Now, we can get new ontology classes from keywords and use them for ontology evolution. Following are steps of new ontology classes recommendation. (1) Get all keywords that appear in subject web-pages, and ignore keywords that are existed in original subject ontology. Finally, a keyword set that contains no duplicate elements is created, and named KeywordSet. (2) For a keyword k_i in KeywordSet, traverse web-pages to get location info of k_i

$$\left\{[n_{title}, \sum_{i=1}^{n_{title}} tf(d_i)], [n_{description}, \sum_{i=1}^{n_{description}} tf(d_i)], [n_{keywors}, \sum_{i=1}^{n_{keywords}} tf(d_i)], [n_{body}, \sum_{i=1}^{n_{body}} tf(d_i)]\right\}. \tag{4}$$

Then, calculate the location weight of k_i with $\omega_{location}$ formula.

(3)For k_i , query the users' behavior logs to get web-pages set named ViewedPageSet that contains web-pages that clicked by users, count totality of users' search named SN , and the frequency that k_i occurs in the search words. Named number of web-pages in ViewedPageSet as CN and traverse ViewedPageSet to count totality of web-pages that contains k_i . Then, calculate the users' behavior weight of k_i with $\omega_{behavior}$ formula .

(4) Combine $\omega_{location}$ with $\omega_{behavior}$ to calculate recommend weight of k_i with $\omega_{recommend}$ formula. (5) Repeat steps (2) to (4) until every keywords in KeywordSet has recommend weight, and a keyword recommend weight set, named Keyword-WeightSet, is created:

$$\{[k_1, \omega_{recommend}(1)], [k_2, \omega_{recommend}(2)], \cdots, [k_n, \omega_{recommend}(n)]\}. \tag{5}$$

(6) For an element t_i in KeywordWeightSet, set a threshold ε , while condition t_i, ε is satisfied, keyword t_i, k is considered to be met requirements of recommendation. (7) Keywords, which are met conditions of recommendation, should be submitted to users, and location of keywords in subject ontology is decided by users.

2.4 Evolution of the domain ontology

Domain ontology consists of classes, relationships and properties. They make up a word set in a hierarchy and could describe subject structurally. There are two main relationships between ontology classes. First, inheritance, a class can only inherited from a parent, and can have several children. The relationships between parent and children is is-a or kind-of. The hierarchy of parent and children could be described as tree. Second, attribute, a class have one or more attributes, relationships between classes and attributes are attribute-of or part-of. In this paper, we choose EMALS (Electromagnetic Aircraft Launch System) as subject, and EMALS is the top class, all the other classer are its sub-classes or attributes. And we choose OWL as development language, and Protg as development environment, and construct EMALS original ontology in Prot$g$ (Fig. 1) by consulting experts, searching relevant literatures, browsing professional books and surfing professing professional websites.

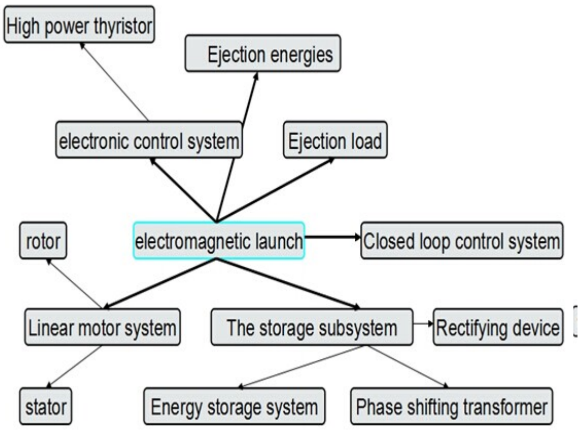


Figure 1: Original EMALS Ontology

After construction of original domain ontology, the main work is to collect subject related web-pages and evolve ontology as mentioned above, and finally a more complete subject ontology is created, as shown in Fig. 2.

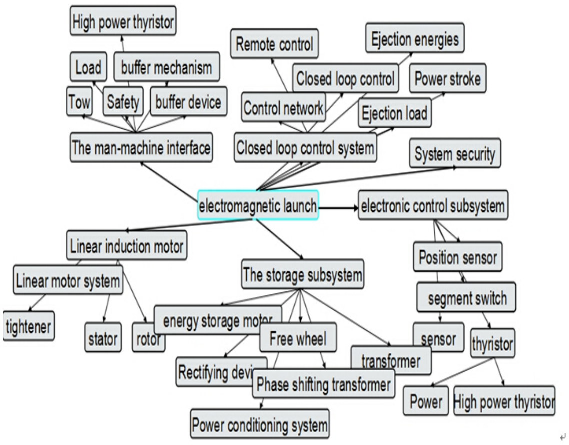


Figure 2: Evolved EMALS Ontology.

3 Experiments and results

3.1 Design of experimentation schema

Subject-relativity of web-page is calculated based on Vector Space model (VSM). First, represent subject-ontology with vector:

$$\alpha = (a_1, a_2, \cdots, a_n), \tag{6}$$

where $a_i = \omega_i$. Then, analysis web-page which is calculating relativity and replace all keywords in web-page with the uniform-description in subject ontology. For example, for keywords K1, K2 and K3, which stands for class C in subject ontology, replace all K1, K2, K3 in web-page to the uniform-description of class C, named CK. Count the frequency of CK. For all frequency of keywords, mark keyword that has max frequency as $k_{f_{max}}$, and named it's frequency as f_{max}. Use f_{max} as a benchmark, and calculate the ratio between other frequency and f_{max}. Set f_{max} as 1, and calculate f_i , which is frequency of keywords. Then weight of keyword is $x_i\omega_i$, and the web-page could be represent as:

$$\beta = (x_1\omega_1, x_2\omega_2, \cdots, x_n\omega_n).$$

Finally, calculate cosine similarity of α and β :

$$\cos(\alpha,\beta) = \frac{(\alpha,\beta)}{|\alpha||\beta|} = \frac{x_1\omega_1^2 + x_2\omega_2^2 + \cdots + x_n\omega_n^2}{\sqrt{\omega_1^2 + \omega_2^2 + \cdots + \omega_n^2}\sqrt{x_1\omega_1^2 + x_2\omega_2^2 + \cdots + x_n\omega_n^2}}. \tag{7}$$

Set a threshold ε , then, web-page is subject related while condition $\cos(\alpha,\beta) > \varepsilon$ was met. Value of ε is depending on research results and practical experience. If much web-pages is needed, web-page subject-relativity should be reduced, and ε is small. While accurate web-page is needed, web-page subject-relativity should be increased, and ε is large. subject ontology evolution proceed is as follows.

(1) Collecting web-pages under guidance of domain ontology;

(2) Parsing and preprocessing subject related web-pages;

(3) Following users behavior of browsing and searching;

(4) Evolving subject ontology by extracting new concepts from web-pages;

Subject related keywords extracted from web-pages and stored in database are assigned recommend weight according to their locations in web-pages and users' behavior tendency. Keywords with high weight will be recommended to users as new subject ontology concepts. And it's users who finally set up relationship between concepts.

(5) Setting up relationships between new concepts and old domain ontology.

3.2 Analysis of experiment results

Table 1: Seed URLs for crawlers

Title	URL
IFeng	http://news.ifeng.com/mil
Sohu	http://mil.sohu.com
Sina	http://mil.news.sina.com.cn
ArmyStar	http://www.armystar.com
Xinhua	http://www.xinhuanet.com/mil

Table 1 shows URL of several portal websites. These URLs are used as seed URLs for crawler to collect subject related web-pages. Crawling each URLs and results show as Fig. 3. From Fig. 3. As we can see, the amount of web-pages crawled increases for some time, and crawling speed is 75 100 pages/min. Then, we try to extract new ontology class from web-pages that have been crawled, and Fig. 4 shows how ε influent the recommendation of new ontology class. Here, ε is a threshold for recommendation weight.

In Fig. 4, we can find that the amount new ontology class recommended is lower while ε is higher, but the result is more accurate and of high subject relativity.

Then, a contrast experiment is conducted between collector with original subject ontology and collector with evolved subject ontology. Fig. 5 is the original ontology of EMALS saved in database. We can see that there is few ontology class. And the collected result for a time under guidance of original ontology is 273 Records. Fig. 6 is the evolved ontology of EMALS saved in database. And it is also used to guide the collection of subject related web-pages, and the result

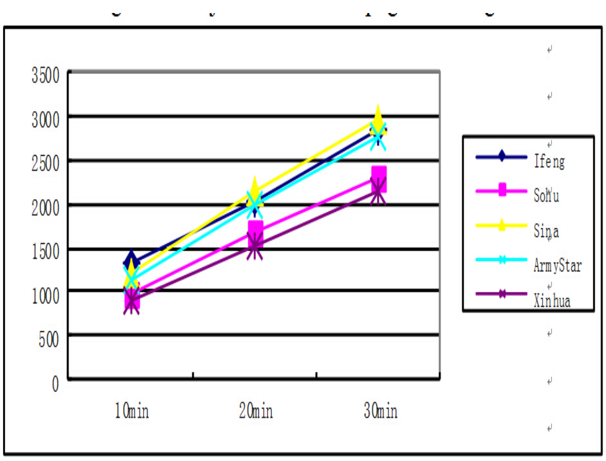


Figure 3: Subject related web-page crawling

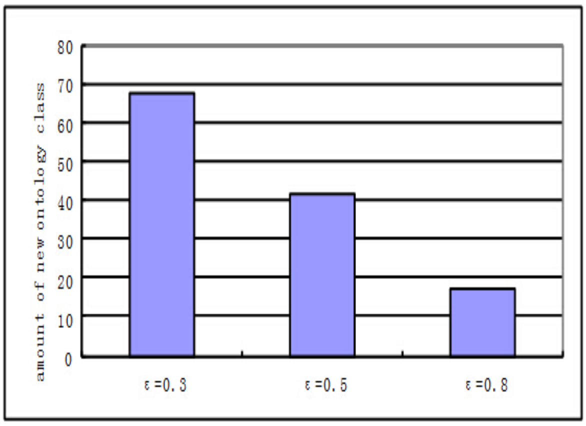
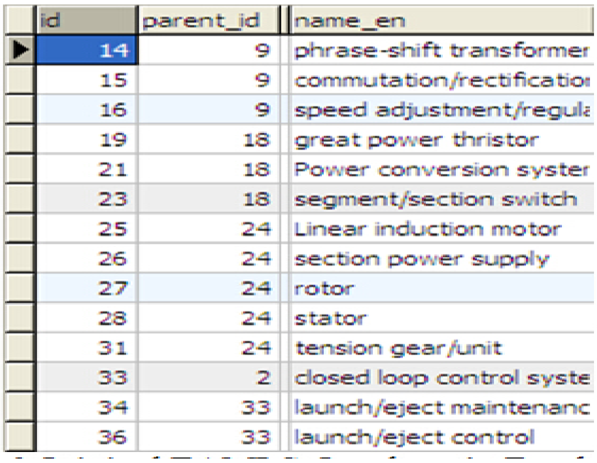
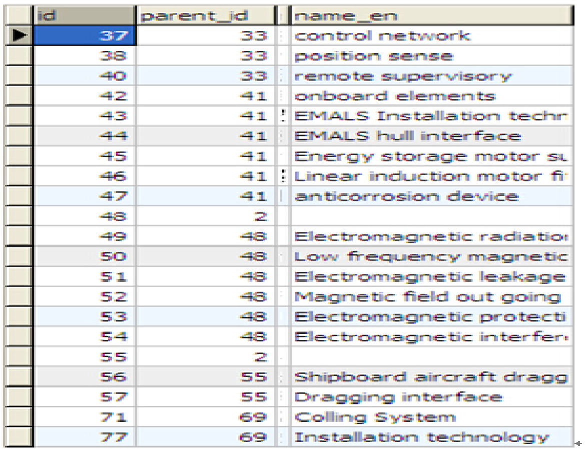


Figure 4: Different ε for new ontology class recommending



id	parent_id	name_en
14	9	phrase-shift transformer
15	9	commutation/rectificatio
16	9	speed adjustment/regula
19	18	great power thristor
21	18	Power conversion syster
23	18	segment/section switch
25	24	Linear induction motor
26	24	section power supply
27	24	rotor
28	24	stator
31	24	tension gear/unit
33	2	closed loop control syste
34	33	launch/eject maintenanc
36	33	launch/eject control

Figure 5: Original EAMLS ontology in database



id	parent_id	name_en
37	33	control network
38	33	position sense
40	33	remote supervisory
42	41	onboard elements
43	41	EMALS Installation techr
44	41	EMALS hull interface
45	41	Energy storage motor su
46	41	Linear induction motor fi
47	41	anticorrosion device
48	2	
49	48	Electromagnetic radiatio
50	48	Low frequency magnetic
51	48	Electromagnetic leakage
52	48	Magnetic field out going
53	48	Electromagnetic protecti
54	48	Electromagnetic interfer
55	2	
56	55	Shipboard aircraft dragg
57	55	Dragging interface
71	69	Colling System
77	69	Installation technology

Figure 6: Evolved EAMLS ontology in database

is 367 Records. Fig. 5 and Fig.6 are the collection experiment results with guidance of original subject ontology and evolved subject ontology. And we can find that during the same time and meeting the same subject relativity, collector with original subject ontology only collects 273 pages while collector with evolved subject ontology collects 367 pages, 34.4% more than the former. This is due to the more complete describe about subject in the evolved ontology, and shows that an gradually evolving ontology contributes to the full-rate of subject information collection.

4 Conclusion

This paper preprocesses subject related web-pages and extracts keywords from description, body etc. Then it calculates location weight based on keywords' position in HTML document and users' behavior weight based on users' behavior factors, such as subject web-pages browsed by user, search words on subject, etc. New Ontology classes are recommended according to their recommend weight, which is calculated by combining location weight with users' behavior weight. The result of experiment shows that this method help to evolve subject ontology well and could improve the efficiency of subject information collection.

Acknowledgments

The authors would like to thank the support from the National Natural Science Foundation of China (Grant Nos. 71373123 and No71271120), and the National Undergraduate Innovation Program of China (Grant No. 2014CX00902).

References

[1] W.F. Ma and X.Y. Du. A Study on Domain Ontology Evolution. *Libary Information Service*, 6(2006):71-74.

[2] D. Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. *Proceedings of the 14th conference on Computational linguistics*, 3(1992):977-981.

[3] M. Sabou, C. Wroe, C. Goble et al. Learning domain ontologies for web service descriptions: an experiment in bioinformatics. *Proceedings of the 14th international conference on World Wide Web ACM*, (2005):190-198.

[4] R. Navigli, P. Velardi and A. Gangemi. Ontology learning and its application to automated terminology translation. *IEEE:Intelligent Systems*, 18(2003):22-31.

[5] M. Zhang, H.T. Geng and X.F. Wang. Automatic Keyword Extraction Algorithm Research Using BC Method. *Journal of Chinese Computer Systems*, 28(2007):189-192.

[6] R.B. Wei. An Empirical Study of Keywords Network Analysis Using Social Network Analysis. *Journal of Intelligence*, 28(2009):46-49.

[7] S.H. Wu and W.L. Hsu. SOAT: a semi-automatic domain ontology acquisition tool from Chinese corpus. *Proceedings of the 19th international conference on Computational linguistics*, 2(2002):1-5.

[8] L. He, H.P. Du and H.Q. Hou. Semi Automatic Construction Method of Domain Ontology. *Library Theory and Practice*, 5(2007):26-27.