An Ontology–Based adaptive Topical Crawling Algorithm

SHEN Jinxing

School of Computer Science ,Guangdong Polytechnic Normal University,China, Guangzhou, 510665

E-mail:jx_shen@sohu.com

Keywords: Ontology, Adaptive, Topical crawling

Abstract: In order to solve the problems of the veracity of the topical web crawler which can not get the information interrelated with the given topical. A design framework and algorithm of the ontology-based adaptive topical crawling is brought forwards. We use the ontology technology to reduce the topical web crawler to get the disrelated information with the topical, in order to improve the correlativity of the topical web crawler. With the experiment of this algorithm, we get the good result of the information with the topical.

Introduction

With the fast evolvement of Internet technology, the Internet has played a more and more critical role in our daily life. World Wide Web (WWW), a technology featured by convenient and straightforward using mode and abundant expression, has witnessed the fastest development and become the most important way to pubic and transmit information. The fast growth of network information resources brings diversified information to us. In this case, the requirement for information is also developing and changing all the time. Now we have fully realized the importance of network information resources. Besides, we have also understood that the true value of information service lies in the content of the provided information. It is not easy to collect information in such a vast information base. Therefore, the search engine [1] technology is introduced. However, current search engines bring some irrelevant information to us. We have to re-judge the returned result and find the required information. Topical Web is then introduced to facilitate a faster and more accurate knowledge acquisition. The topical web administrator, however, faces the challenge to collect the information for the topical web, which cannot be done manually. In this case, the crawler based on ontology is offered. In the topical crawler, the core is the crawling algorism [2]. Topical Web crawling, put it in another word, Focused Web Crawling is the key technology to acquire pages of specific topics in World Wide Web (WWW). When information in the Internet grows exponentially, the specialized search engines for certain topics become even more applicable, and can serve as the supplement for universal search engines.

Based on the traditional search engine technology, Topical Web Crawling applies technologies including ontology, text classification, cluster, and web mining to capture the page information of specific topics. In this way, the search accuracy can be improved, the occupation of network resources can be reduces, and the update period for the data of pages can be shortened.

The essence to construct specialized search engine is to search web pages relevant to a specific topic. This paper proposes a new self-adaptive crawling algorithm based on ontology. The experiment shows a good result.

Ontology and Topical Web

Ontology is originally a topic in Philosophy, which generally describes the ontology in the real world, and serves as the general theory of ontology existence and ontology hypostases [7]. After being introduced to the area of artificial intelligence, ontology has a more specific meaning. Because people did not have a thorough understanding of ontology at the beginning, the definition of this concept is evolving all the time. At present, the definition of ontology in artificial intelligent

© 2015. The authors - Published by Atlantis Press

[8] is well accepted, which can be summarized as follow: Ontology is definite and formalized specification of shared concept model. It indicates four aspects: concept model, definitiveness, formalization, and sharing. Concept model is a kind of model extracted from the phenomenon in the real world, but not the same as that in the real world. Concept model is something like a dictionary or a glossary, which is composed of concepts, axioms, and relations. Definitiveness means that all the concepts and concept restrictions have clear definitions and explanations. Formalization indicates that the content in ontology can be understood and processed by the computer. Sharing shows that the knowledge in ontology is widely accepted, rather than private and accepted by only a certain group. In this way, the knowledge sharing and knowledge re-application can be achieved. Sharing and re-application is two important advantages of ontology.

Topical web is a cluster of specialized information. Information of each topic is an integration of relevant information in several websites. Information in a topical web is received from a knowledge base, which in turn finds the information in each website. Therefore, the information of knowledge base depends on information of each website.

Ontology specifies the exact meaning of a concept by the strict definition of the concept and the relationship between concepts. In this way, ontology can indicates the well acknowledged and shared knowledge. Therefore, a group of concepts and concept relations can be summarized according to the specific area in the real world. Then the ontology can be constructed, which can facilitate information processing for this area. Topical Crawler Algorithm Based on Ontology ensures high crawling efficiency.

Framework of Self-adaptive Knowledge Topical Crawler Based on Ontology

A. Basic Principle of Network Crawler

The search engine has the following working process [3]:

1. Find and collect webpage information on the Internet.
2. Sort and organize the information to establish an index.
3. Quickly find the document in the index and sort the output according to the key word input by the user. Then return the result to the user.

To find and collect webpage information, there should be a crawler program with high performance. The following is a typical crawler working mode: check a webpage and find the relevant information. Then continue the search process through all the links in this webpage [4].

The specific working mode of a crawler:

In most cases, a crawler algorithm carries out the following:

1. Take a URL from a URL list.
2. Analysis the host addresses of this URL and download the relevant documents.
3. Analysis these documents and get the new URL.
4. Find the relative address according to the newly analyzed link.
5. Add the new URL into the URL list if it is met for the first time.

That is to say, the basic algorithm has the following:

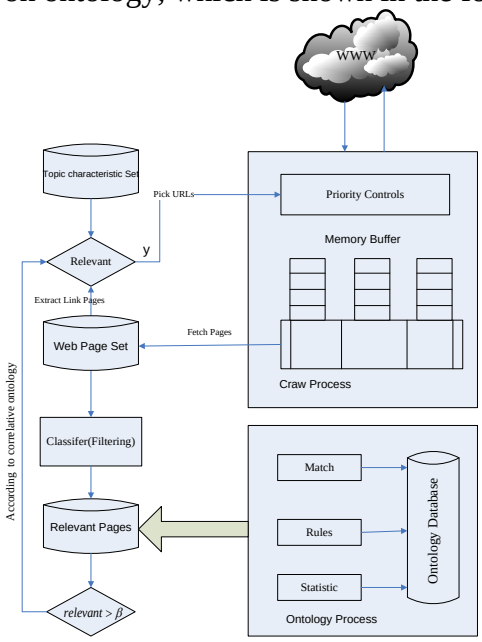
1. A part is in charge of storing the URL to be downloaded.
2. A part is in charge of converse the domain name of the host into the corresponding IP address.
3. A part is in charge of downloading files in HTTP protocol.
4. A part is in charge of analyzing the new URL from the HTML file.
5. A part is in charge of checking whether the URL is ever met.

B. Basic Framework of Self-adaptive Knowledge Topical Crawler

The prototype of knowledge crawler is composed of user interface, pre-processing, knowledge model [5], crawler, knowledge processing and so on. The input sub-system requires that the user inputs the parameters such as key concept of the required knowledge [6] (get the URLs of the seed), the queried knowledge, and layer. On the other hand, the output sub-system outputs URLs sorted by the knowledge and knowledge segments. The pre-processing component refers to how to effectively find the relevant web pages and sorts the URLs. Firstly select and download the web-pages

according to the principle of the most similar path and semantic structure of subject anchor. Then segment these web pages according to the content of the title, and find the links of relevant links. Re-process these links and analyses the anchor text to facilitate URL sorting.

The webpage search refers to algorithm to find the relevant web pages based on the principle of the most relevance. This is actually a graphic search issue. It is critical to select the seed webpage when knowledge acquirement, which can be done by Google API or by the suggestion from the expert. As for the mode to process the acquired seed, this paper proposes a new crawler framework after a comprehensive study of current crawler frameworks, that is the framework of self-adaptive knowledge topical crawler based on ontology, which is shown in the following figure:



C. An Ontology-Based adaptive Crawling Algorithm

In the frame of ontology-based intellectively adaptive Topical Web Crawling which described above, the most important part is Crawling algorithm. In the process of Topical Web Crawling, an adaptive way of ontology-base is used, which uses excellent and relevant ontology inter linkage to consummate the topic characteristic set according to context information. After crawling program downloaded the objective web page to local from crawling queue, the topic correlation degree was estimated by a universal categorizer. If the web page was relevant enough, its URL inter linkage was interlinked the context to extract the character and added into topic characteristic set to enrich the topic characteristic set. When web page had been estimated, all interlinked context was picked and computed its priority. Then un-crawling URLs was added to crawling queue and sorted by their priority. The inter linkage priority $R(u)$ is given as:

$$R(u) = \alpha * R(page) + (1-\alpha) * sim(c,q),\ 0 < \alpha < 1$$

Where, α is a parameter, which is used to adjust the specific weight between the topic correlation degree of web page and the topic correlation degree of interlink age context. As the choice of the value of α, it is given in next section.

Input: Seed URLs queue: $starting\_url$

Algorithm begin

1. *for each u in Seed URLs*
2. $DB = fetch\_context(u, N)$ //extract N link context to each seed URL
3. $featureSet = extract\_features(DB)$ //extract topic characteristic set
4. $enqueue(url\_queue, starting\_url)$ // put staring URL into crawling queue
5. $while(|url\_queue| > 0)$

6. $url = dequeue(url\_queue)$
7. $page = crawl\_page(url)$ //get the web page linked by inter linkage url
8. $enqueue(crawled\_queue, url)$ // put the url into crawled queue
9. $if\ relevance(page) \geq \beta$ // page is relevant enough
10. $relevant\_pageDB = page$
11. $goodlinkcontextDB = extract\_linkcontext(url)$ //extract the url context
12. $if\ goodlinkcontextDB \geq \lambda$ // when $goodlinkcontextDB$ is bigger than a value define before
13. $featureSet = featureSet + extract\_features(goodlinkcontextDB)$
14. $link\_contexts = extract\_contexts(page)$ //extract link context of web page
15. $evaluate(link\_contexts)$ //compute the correlation of each link context, and evaluate priority of each url
16. $url\_list = extract\_urls(page)$
17. $for\ each\ u\ in\ url\_list$
18. $if\ (u \notin url\_queue\ and\ u \notin crawled\_queue)$
19. $enqueue(url\_queue, u)$
20. $reorder\_queue(url\_queue)$

End.

Attention:

$enqueue(queue, element)$: add element into queue

$dequeue(queue)$: return the front element in queue and remove it out of queue

$extract\_linkcontext(url)$: extract link context of inter linkage URL

$extract\_contexts(page)$: extract all link contexts of URL in web page

$recorder\_queue(queue)$: Recorder $queue$ according to the priority

$extract\_features(db)$: extract optic characteristic set from db

Performance analysis

For the algorithm, experiment was implemented on key word-based topic crawling and ontology-based adaptive crawling. Software and hardware environment is constant： We use the computer developed in JBuilder2007, CPU: P4 2.4G， EMS memory: 1G. In order to compare with the topic crawling without ontology, the experiment is around the subject of indicator, because information valuating system mostly evaluates the information about monitor at present. The parameter used in this experiment is: searching depth=2(designed a little small in case searching scope is too big), No. thread= 200(used in good network environment), initial seed=10 (good seed by selection), thread r=0.1

Two kinds of change were found after comparing before-and-after imported date:

1) The number of extracted documents was reduced. Link analyzed and deleted a great deal of irrelevant web page and few web pages whose degree of correlation was prodigious was also deleted, according to carefully comparing. These links mostly were up-link, crosslink and extroversion-link. Introduced ontology made the number of relevant web page decrease.

2) Crawling time was reduced. After a great deal of irrelevant web page was deleted, crawling lode was reduced.

types	Key word-based		Ontology-based	
	before link analysis	after link analysis	before link analysis	after link analysis
Number of extracted documents	2987	2107	2315	1126
No. failing documents	587	189	132	21
No. refused documents	1	0	0	0
No. found documents	3516	2317	2523	1154
Collected date /Bytes	94583726	63591475	71381957	43625919
Crawling time/s	364	259	584	469

In conclusion, after linkage analyze of ontology was introduced, crawling precision was not affected but crawling rate was increased. Therefore, this algorithm is successful in the frame of above mention.

Conclusions

Human can forecast objective web page is weather for what they need via their knowledge structure and environments information. In order to make computers forecast topic correlation of objective web page using web link information, they must have abundant knowledge and prompt information. As the use of web environment information, computer cannot forecast objective web content topic correctly from the link anchor information like human. In order to let computer forecast topic correlation of pointed objective web page by linkage, it's not enough just from the link anchor information. We must increase or filter more abundant and exact forecast information to decide topic correlation of objective web page. Therefore, this section analyses the structure of web page and extracts linkage context by a method of ontology-based to extend anchor context information appropriately. Then we use this information to evaluate the topic correlation of the web page pointed by linkage. This paper proposed an ontology-based adaptive crawling algorithm which evaluates the topic correlation of web page by using ontology categorizer and extracts more topic character to enrich characteristic set. The experiment shows that this algorithm improved the topic crawling performance, meanwhile, topic shift was not found. An ontology-based adaptive crawling algorithm needs further research, such as, the research about the improvement of its performance, the research on the improvement of self – ratiocination in the case that ontology database is not consummate.

Reference

[1] Boris Rotenberg, Ramón Compañó,Search Engines for Audio-Visual Content: Copyright Law and Its Policy Relevance. Telecommunication Markets Contributions to Economics 2009, pp 113-139,Feb 2009.

[2] Ari Pirkola, Tuomas Talvensaari,A Topic-Specific Web Search System Focusing on Quality Pages.Research and Advanced Technology for Digital Libraries,Lecture Notes in Computer Science Volume 6273, 2010, pp 490-493

[3] M. Ehrig, A. Maedche. Ontology-focused Crawling of Web Documents, In Proceedings of the 2003 ACM symposium on Applied computing, 2003

[4] Menczer F, Pant G, Srinivasan P.Topic Web crawlers: Evaluation adaptive alogorithms .ACM Trans. on Internet Technologies 2003 26(1):89~113
[5] TANG Jie,LIANG Bang-yong,LI Juan-zi,WANG Ke-long. Automatic Ontology Mapping in Semantic Web[J]. Chinese Journal of Computers. Vol.29,No.11,2006.
[6] AnHai Doan, Jayant Madhavan, Pedro Domingos,et Learning to map between ontologies on the semantic web[A].Proceedings of the 11th international conference on world wide web [C].2002.662-673.
[7] Maedche A, Stab S.Measuring similarity between ontologies [A],Proceeding of the European conference on knowledge acquisition and management[C].2002
[8] M Uschold. Ontologies principles , methods and applications[A] . Knowledge Engineering Review[C] . Boston , US: IEEE Press , 1996. 213 ~ 228.