Semantic Focused Crawler Using Ontology in Web Mining for Measuring Concept Similarity

[1]N.Vijayalakshmi, [2]J.Sangeetha

[1]Assistant Professor, [2]M.Phil Scholar
Department of Computer Science
Kovai Kalaimagal College of Arts and Science,Coimbatore, India

Abstract— **In the world of Internet, semantic crawlers played a vital role in optimizing the user query search in Web Data Mining. An unsupervised ontology learning algorithm is used in self adaptive semantic crawlers to maintain the performance of the crawlers. For each concept from the crawled web page, a value is calculated, which is used to train the values of the concept in later search. The dynamic nature of concept values and learned description values has been a reason in performance declination. In the proposed work the usage of JCN algorithm computes the semantic relatedness of word senses. It measures the edge counts using a 'is-a' hierarchy and Information Content Values in WordNet. Ontology learning and data mining techniques is used for effective evaluation measures, which can be used to select an ontology that is best out of many candidates. The implemented algorithm incorporates the technologies of semantic focused crawling and ontology learning, in order to maintain the performance of the crawler in Web Mining, regardless of the variety in the Web environment. The innovations of the algorithm is based on the design of an unsupervised framework for vocabulary-based text data mining , and a Jiang Conarth algorithm for matching semantically relevant concepts.**

INTRODUCTION

Web services provide access to software systems over the Internet using standard protocols. In the most basic scenario there is a Web Service Provider who publishes a service and a Web Service Consumer who uses this service. Web Service Discovery is the process of finding a suitable Web Service for a given task. Publishing a Web service involves creating a software artifact and making it accessible to potential consumers. Web Service Providers augment a Web service endpoint with an interface description using the Web Services Description Language so that a consumer can use the service. Optionally, a provider can explicitly register a service with a Web Services Registry such as Universal Description Discovery and Integration (UDDI) or publish additional documents intended to facilitate discovery such as Web Services Inspection Language (WSIL) documents. The service users or consumers can search Web Services manually or automatically.

Web Mining

Web mining is the application of data mining techniques to Web data. Web mining helps to solve the problem of discovering how users are using Web sites. It involves log analysis (or log analysis) and the steps that typically have to be gone through to get meaningful data from Web logs like data collection, pre-processing, data enrichment and pattern analysis and discovery. Web mining describes the application of traditional data mining techniques onto the web resources and has facilitated the further development of the techniques to consider the specific structures of web data. The analyzed web resources contain (1) the actual web site (2) the hyperlinks connecting these sites and (3) the path that online users take on the web to reach a particular site.

Web Usage Mining

Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications.

Web Structure Mining

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds: (1) Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location. (2)Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, ALIWEB,Meta Crawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. In recent years these

factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the web.

Ontology

Ontologies constitute a formal conceptualization of a particular domain of interest that is shared by a group of people. When building ontologies into information systems, it is possible to modularize many software aspects mostly related to the domain (e.g., taxonomic structures) from ones mostly related to the processing (e.g., querying) and visualization of data. One could argue that the drawback one encounters there is that such information systems software cannot be built with an implicit understanding of the domain, but rather it is necessary to make conceptualizations of the domain explicit — which may be a difficult task, resulting in a well-known knowledge engineering bottleneck.

While one answer to this argument, also found in software engineering, certainly is: you should make your structures explicit in order to be able to adapt and extend them easily, the quest for faster and cheaper ontology engineering remains. Though ontology engineering tools have matured over the last decade, the manual building of ontologies still remains a tedious, cumbersome task. Ontology Learning aims at the integration of a multitude of disciplines in order to facilitate the construction of ontologies, in particular ontology engineering and machine learning. Because the fully automatic acquisition of knowledge by machines remains in the distant future, the overall process is considered to be semi-automatic with human intervention.

Crawlers

Semantic crawlers are a variation of classic focused crawlers. To compute topic to page relevance downloaded priorities are assigned to pages by applying semantic similarity criteria: the sharing of conceptually similar terms defines the relevance of a page and the topic. Ontology is used to define the conceptual similarity between the terms. Learning crawlers uses a training process to guide the crawling process and to assign visit priorities to web pages. A learning crawler supplies a training set which consist of relevant and not relevant Web pages in order to train the learning crawler. Links are extracted from web pages by assigning the higher visit priorities to classify relevant topic. Methods based on context graphs and Hidden Markov Models (HMM) take into account not only the page content but also the link structure of the Web and the probability that a given page (which may be not relevant to the topic) will lead to a relevant page.

Architecture

A crawler must not only have a good crawling strategy, as noted in the previous sections, but it should also have a highly optimized architecture. It is fairly easy to build a slow crawler that downloads a few pages per second for a short period of time, building a high-performance system that can download hundreds of millions of pages over several weeks presents a number of challenges in system design, I/O and network efficiency, and robustness and manageability. Web crawlers are a central part of search engines, and details on their algorithms and architecture are kept as business secrets. When crawler designs are published, there is often an important lack of detail that prevents others from reproducing the work. There are also emerging concerns about "search engine spamming", which prevent major search engines from publishing their ranking algorithms.Web crawlers typically identify themselves to a Web server by using the User-agent field of an HTTP request. Web site administrators typically examine their Web servers log and use the user agent field to determine which crawlers have visited the web server and how often. The user agent field may include a URL where the Web site administrator may find out more information about the crawler. Examining Web server log is tedious task therefore some administrators use tools such as CrawlTrack or SEO Crawlytics to identify, track and verify Web crawlers. Spambots and other malicious Web crawlers are unlikely to place identifying information in the user agent field, or they may mask their identity as a browser or other well-known crawler.It is important for Web crawlers to identify themselves so that Web site administrators can contact the owner if needed. In some cases, crawlers may be accidentally trapped in a crawler trap or they may be overloading a Web server with requests, and the owner needs to stop the crawler. A vast amount of web pages lie in the deep or invisible web.

PROCESS OF SEMANTIC FOCUSED CRAWLER USING ONTOLOGY IN WEB MINING FOR MEASURING CONCEPT SIMILARITY

Preprocessing

Preprocessing which is to process the contents of the concept description property of each concept in the ontology before matching the metadata and the concepts. The documents are prepared for the extraction. A complete methodology for automatic knowledge extraction, in the form of ontological concepts, from a knowledge base of heterogeneous documents. The documents are converted from the original format to a more suitable one.

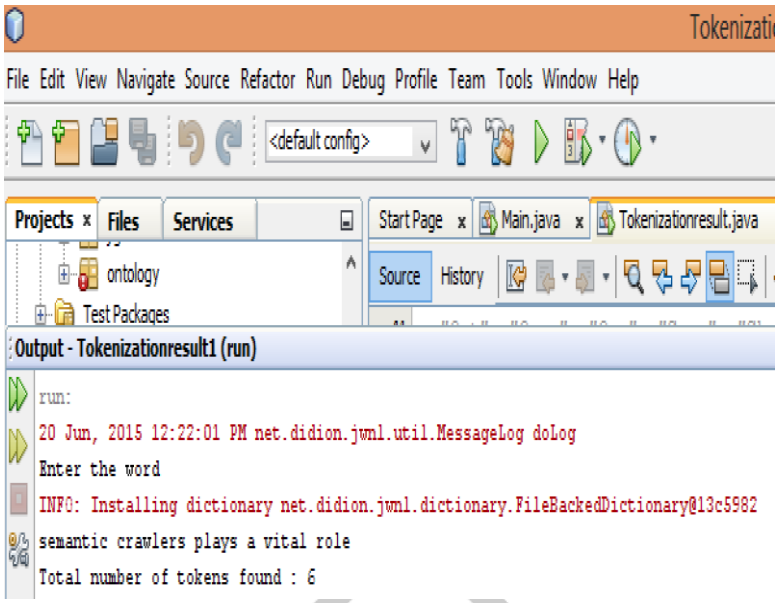

Figure 1 Preprocessing of Web Pages content

POS Classification

It represents the process of marking the terms in the document (including terms composed of several words) in a text as corresponding to a particular part of speech (i.e., names, verbs, adjectives, adverbs, etc.).

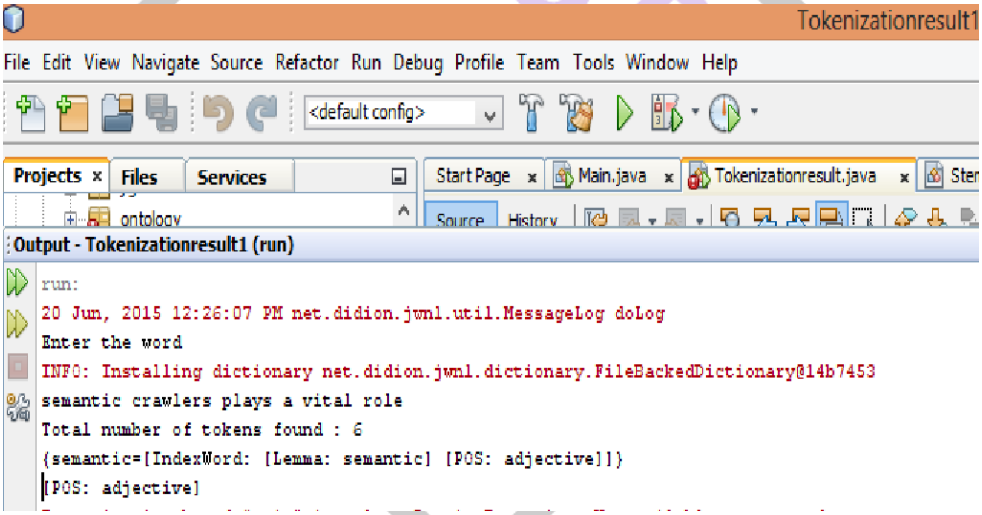

Figure: 2 POS Classification

Stemming

It is the process of reducing a term of the analyzed document to its stem or root form (e.g., writing → write). The stem does not need to be identical to the morphological root of the term; it is usually sufficient that related words map to the same stem, even if this stem is not a valid root.

Synonyms Searching

The WordNet lexical database for the acquisition of the synonyms of a term: The acquired terms are associated to the first term and are taken into account during the text processing.

Crawling and Term Extraction

Two processes crawling and extraction is to download Web pages from the Internet at one time, and to extract the required information from the downloaded Web pages, according to the mining service metadata schema and the mining service provider metadata schema in order to prepare the property values to generate a new group of metadata. These two processes are realized by the semantic focused crawler. The next step is term processing, which is to process the content of the service Description property of the metadata in order to prepare for subsequent concept-metadata matching. A simple draft version of the ontology is created. From the syntactic point of view, we assume that the elements of interest for the user are constructed on the grounds of some primitive terms. According to this assumption, ontology consists of primitive classes and compound classes.

JCN Algorithm

This module computes the semantic relatedness of word senses. This measure is based on a combination of using edge counts in the WordNet 'is-a' hierarchy and using the information content values of the WordNet concepts. Their measure in mining web , however, computes values that indicate the semantic distance between words .In this implementation of the measure we invert the value so as to obtain a measure of semantic relatedness. Other issues that arise due to this have been taken care of as special cases.

Estimation of Semantic Relevance

If concept description property and learned concept description property holds the service description property then it is semantically relevant. This is estimated using string matching in JCN algorithm. This helps in generating the metadata and metadata association and can be stored in metadata base.

Algorithms

Concept Metadata Semantic Algorithm

This is implemented by combining the Semantic based and Statistics based String Matching Algorithm. The former one measures the text similarity between a concept and service description and uses Resnik's model.

Categorization Algorithm

A categorization algorithm is chosen which accepts learned concept description property value and concept description property value and helps in training the classification of ontology metadata base. The algorithm called Winnowing uses a multiplicative weight update score and is capable of performing much better when too many data are irrelevant. It also scales well on high dimensional data.

String Matching Algorithm

Semi supervised ontology method automatically obtains the statistical data from the Web pages, in order to compute the semantic relevance between a service description and a concept description of a concept. In Statistics-Based String Matching Algorithm follows semi- supervised training paradigm aimed to finding the maximum probability semantic relevance and co-occur in the Web pages. Automated ontology learning techniques also require effective evaluation measures, which can be used to select the best ontology out of many candidates, to select values of tunable parameters of the learning algorithm, or to direct the learning process itself if the latter is formulated as finding a path through a search space.

Semantic-Based String Matching Algorithm

The key idea of the SeSM algorithm is to measure the text similarity between a concept description and a service description, by means of WordNet9 and a semantic similarity model. As the concept description and the service description can be regarded as two groups of terms after the preprocessing and term processing phase, first of all, we need to examine the semantic similarity between any two terms from these two groups. Here we make use of Resnik's information-theoretic model and Word Net to achieve this goal. Since terms (or concepts) in Word Net are organized in a hierarchical structure, in which concepts have the relationships of hypernym/hyponym, it is possible to assess the similarity between two concepts by comparing their relative position in Word Net.

Statistics-Based String Matching Algorithm

A statistics-based model to achieve this goal. In the crawling process and the subsequent processes the SASF crawler downloads K Web pages at the beginning, and automatically obtains the statistical data from the Web pages, in order to compute the semantic relevance between a service description (SDi) and a concept description (SDjh) of a concept (Cj). The StSM algorithm follows semi- supervised training paradigm aimed to finding the maximum probability semantic relevance and co-occur in the Web pages and automated ontology learning techniques also require effective evaluation measures.

Performance Measure

Table 1 Comparison of Harvest Rate of Five Crawlers

S.NO	METHODS	HARVEST RATE
1	Breadth First Crawler	0.546
2	SSRM Crawler	0.604
3	VSM Crawler	0.661
4	CMCFC	0.763
5	SASF Crawler	0.6

III. CONCLUSION AND FUTURE WORK

Most of the ontology focused crawlers in web data mining have the limitation that they could not able to evolve ontologies by enhancing the vocabulary collections. A supervised ontology learning crawler enhances the harvest rate of crawling without considering the classification. It may not even work in an uncontrolled web environment when new unpredicted term appears. This leads to the usage of Ontology learning based focused crawlers. By considering the three issues heterogeneity, ubiquity and ambiguity in the field of Service Data Mining, an adaptive focused crawler was developed. In Self Adaptive Semantic Focused Crawler it follows unsupervised learning framework in ontology learning and a concept-metadata matching algorithm is used for

finding relevance between service concept and service metadata. In this approach the algorithm based string matching process which cannot accept learned concept description property values. This can be solved by involving categorization algorithms such as JCN Algorithm which accepts learned concept description property values and also helps in determining the boundary values based on appropriate categorization algorithms. With the implementation of the JCN algorithm an improvement in accuracy of semantic relatedness is achieved. In future work a semi supervised algorithm can be used to improve the accuracy of the semantic relevance of concepts in Web Mining. The semi supervised algorithm is a combination of supervised and unsupervised algorithm.

REFERENCES

[1] Alessandro Micarelli and Fabio Gasparetti Adaptive Focused Crawling Department of Computer Science and Automation Artificial Intelligence Laboratory Roma Tre University

[2] S. Castano, A. Ferrara, S. Montanelli, G. Racca (2004), 'From Surface to Intensive Matching of Semantic Web Ontologies'. IEEE Database and Expert System Applications. Pages 140-144.

[3] Charlotte Lecluze, Lo ¨ ıs Rigouste, Emmanuel Giguet, Nadine Lucas (2013), 'Which Granularity to Bootstrap a Multilingual Method of Document Alignment: Character N-grams or Word N-grams?', Procedia - Social and Behavioral Sciences 95 (2013) 473 – 481 .

[4] Charlie Greenbacker , 'WordNet Similarity Metrics', Speech and Language Processing (2nd Ed.) Ch. 20.6 Word Similarity: Thesaurus Methods, plus NLTK documentation (PPT).

[5] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, Member, IEEE (2011), 'A Web Search Engine-Based Approach to Measure Semantic Similarity between Words'. IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 7.

[6] Debajyoti Mukhopadhyay, Archana Chougule(2013), 'A Framework for Semi-automated Web Service Composition in Semantic Web'. CUBE International Conference.

[7] I. M. Delamer and J. L. M. Lastra, "Service-oriented architecture for distributed publish/subscribe middleware in electronics production," IEEE Trans. Ind. Informat., vol. 2, no. 4, pp. 281–294, Nov. 2006.

[8] Esko Ukkonen (1985), 'Algorithms for Approximate String Matching', Information And Control 64, 100--118.

[9] Esko Ukkonen, (1992) 'Approximate string-matching with q-grams and maximal matches', Theoretical Computer Science 92 .191-211 Elsevier

[10] Gang Liu1,Ruili Wang1,Jeremy Buckley,Helen M. Zhou (2011), 'A WordNet-based Semantic Similarity Measure Enhanced by Internet-based Knowledge' .23rd International Conference on Software Engineering and Knowledge Engineering. Pages 175-178.

[11] Gheorghe Pa˘un (2010), 'A quick introduction to membrane computing', The Journal of Logic and Algebraic Programming Vol. 79 , Pages 291–294.

[12] Giuseppe Fenza,Vincenzo Loia, Sabrine Senatore (2008), 'A Hybrid Approach to Semantic Web Services Match Making ', International Journal of Appropriate Reasoning, Vol.48,Pages 808-828

[13] Gonzalo Navarroa, Edgar Chávezb, 'A metric index for approximate string matching ', Theoretical Computer Science 352 .266–279. 2006.

[14] T. R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, vol. 5, pp. 199–220, 1993.

[15] Hai Dong, Member, IEEE, and Farookh Khadeer Hussain (2014), 'Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery', IEEE Transactions on Industrial Informatics, Vol. 10, no. 2.

[16] Harry T Yani Achsana, Wahyu Catur Wibowob (2014), 'A Fast Distributed Focused-Web Crawling', Elsevier Ltd. Selection and peer-review under responsibility of DAAAM International Vienna doi, Proceeding Engineering 49,Pages 492-499.

[17] Imants Zaremboa, Artis Teilansa, Aldis Rausisb, Jazeps Bulsb (2015), 'Assessment of Name Based Algorithms for Land Administration Ontology Matching', Procedia Computer Science 43 Pg 53 – 61 .

[18] Jaytrilok Choudhary, Devshri Roy (2013), 'Priority based Semantic Web Crawler', International Journal of Computer Applications ,Vol. 81, No 15.

[19] Jike Ge, Yuhui Qiu(2008), 'Concept Similarity Matching Based on Semantic Distance', Fourth International Conference on Semantics, Knowledge and Grid, Pages 380-383

[20] Lefteris Kozanidis (2008), 'An Ontology-Based Focused Crawler ',NLDB '08 Proceedings of the 13th international conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems, Pages 376 – 379.

[21] Morteza Okhovvat ,Behrouz Minaei Bidgolib (2011), 'A Hidden Markov Model for Persian Part-of-Speech Tagging ', Procedia Computer Science 3. Pg no: 977-981.

[22] Markus E. Nebel (2006), 'Fast string matching by using probabilities: On an optimal mismatch variant of Horspool's algorithm',Theoretical Computer Science 359.329–343.

[23] Ngo Xuan Bacha,, Kunihiko Hiraishia, Nguyen Le Minha, Akira Shimazua(2013), 'Dual Decomposition for Vietnamese Part-of-Speech Tagging'. Procedia Computer Science 22 .Pg 123 – 131

[24] Nidhi Jain, Paramjeet Rawat (2013), 'A Study of Focused Web Crawlers for Semantic Web', International Journal of Computer Science and Information Technologies, Vol. 4, No 2,Pages 398-402.

[25] Parvaneh Khosravizadeha, Roya Pashmforoosha (2011), 'How parts of speech are learned? A lexical-driven or a structure-driven model ',4th International Conference of Cognitive Science (ICCS 2011).

[26] M. Ruta, F. Scioscia, E. Di Sciascio, and G. Loseto, "Semantic-based enhancement of ISO/IEC 14543–3 EIB/KNX standard for building automation," IEEE Trans. Ind. Informat., vol. 7, no. 4, pp. 731–739, Nov. 2011.

[27] Raphael Clifford, Benjamin Sach (2011), 'Pattern Matching in Pseudo RealTime', Journal of Discrete Algorithms 9. Pg 67-89.