ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Clustered Based User-Interest Ontology Construction for Selecting Seed URLs of Focused Crawler

J. Nisha[1], K. Sundareswari [2]

III M.E., Dept of CSE, Karpagam University, Coimbatore , India[1]

Assistant Professor , Dept of CSE, Karpagam University, Coimbatore , India[2]

ABSTRACT*:* With the increasing number of accessible web pages on Internet, it has become gradually difficult for users to find the web pages that are relevant to their particular needs. Knowledge about computer users is very beneficial for assisting them, predicting their future actions. Seed URLs selection for focused Web crawler intends to guide related and valuable information that meets a user's personal information requirement and provide more effective information retrieval. In this paper, a seed URLs selection approach is proposed based on user-interest ontology. In order to enrich semantic query, first intend to apply Formal Concept Analysis to construct user-interest concept lattice with user log profile. By using concept lattice merger, construct the user-interest ontology which can describe the implicit concepts and relationships between them more appropriately for semantic representation and query match. On the other hand, make full use of the user-interest ontology for extracting the user interest topic area and expanding user queries to receive the most related pages as seed URLs, which is an entrance of the focused crawler. In particular, focus on how to refine the user topic area using the bipartite directed graph.

KEYWORDS: URL , data mining , crawler , ontology

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed. That particular data may come from all parts of business, from the production to the management. Managers also use data mining to decide upon marketing strategies for their product. They can use data to compare and contrast among competitors. Data mining interprets its data into real time analysis that can be used to increase sales, promote new product, or delete product that is not value-added to the company.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

II. RELATED WORK

[1] Topic-specific web crawler is a program used for searching information related to some interested topics from the Internet. The main property of topic-specific crawling is that the crawler does not need to collect all web pages, but selects and retrieves relevant pages only. Because the crawler is only a computer program, it cannot determine how relevant a web page is. The rapid growth of the Internet has put us into trouble when we need to find information in such a large network of databases . Authors present an algorithm that covers detail of the first and the consecutive crawling. We add a learning ability from previous crawling to improve the efficiency of consecutive crawling processes. In [2] authors propose a method to build a dynamic representation of the semantic context of ongoing retrieval tasks, which is used to activate different subsets of user interests at runtime, in a way that out–of-context preferences are discarded. This approach is based on an ontology driven representation of the domain of discourse, providing enriched descriptions of the semantics involved in retrieval actions and preferences, and enabling the definition of effective means to relate preferences and context. [3] Because of the brevity and semantic ambiguity of user query words, most search engines face a problem to understand the meaning of query words. topic search engine to not only accurately understand user submitting information needs, but also possess of the relevant semantic knowledge of query information source, and how to automatically and distinguishingly return the accurate relevant information to each user when "different users enter the same query keywords" and "the same user inputs different query keywords" to topic search engine, which is our main research issues. Author proposes an approach to construct user query ontology based on WordNet and the query clustering results. User query ontology can specific an interest area for a given query terms, which form the basis of personalized intelligent information retrieval. In [4] authors provide a mechanism for temporally tracking down changes to an ontology throughout it's life span. In particular the objective is to work on ontology change management, recovery, and visualization of changes and their effects on ontology to understand the ontology's evolution behavior. To achieve this, all these changes are maintained and managed in a coherent manner. A Semantically enriched Change History Ontology (CHO) is developed and used to record the ontology changes in a Change History Log (CHL). For proof of the concept, here developed the system as a plug-in for the ontology editor Protege that listens and logs all of the ontology changes in CHL. Afterwards, these logged changes are used for ontology recovery (Roll Back and Roll Forward) purposes. Here designed and implemented the Roll Back and Roll Forward algorithms. The logged changes are also used for visualization of changes and their effects at different stages of the evolving ontology. A play back feature is provided to navigate the ontology history for a better understanding of its evolution behavior. In [7] authors updated CCG based on incremental learning to get more topic relevant web pages. Author extracted some Incremental Concept (IC) from new visited pages and inserted these IC into CCG by the semantic similarity between core concept and incremental concept. In addition, authors deleted some concepts from CCG according to a given threshold . lastly, the experiment proved that there was a better result in focused web crawling by our method.

III. PROPOSED SYSTEM

A. *Description of the Proposed System:*

A novel approach for selecting seed URLs of the web focused crawler based on the user interest ontology is proposed. In the search engine, it collects an amount of history behaviors which the user visited the search result and forms user log profile. Every record in user log profile contains user query topics and the corresponding clicked URLs. This history information of the records hides rich knowledge reflecting word-to-word and concept-to-concept in user interest domain. User interest ontology is constructed using user knowledge.

When a user submits his query topics to our Personalized Intelligence Search Engine (PISE), the PISE expands user query topics by user-interest ontology and calls other general search engines, such as Google, Yahoo and AltaVista, etc. Some web pages and their URLs are returned. These web pages can be vectored. Fully consider the semantic relationship of words (or concepts) in user interest ontology to optimize these vectors. All web pages are considered to construct user-interest vector in user log profile. From that common words are selected.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

User profile creation

If any new users enter into the system, he/she should register with their details first before login. This module is for getting the query from the user. And if the user gives the query in the search, each query is noted with the particular search time and it stored in the database. The current user is identified by using session attributes.

Creating user Interest Ontology

In this module user interest ontology is created by using the user history query logs. User entered queries are typically found within the query logs of a search engine. If two queries that are issued consecutively by many users occur frequently enough, they are likely to be reformulations of each other. To measure the relevance between two queries issued by a user, makes use of the interval between the timestamps of the queries within the user's search history.

Ranking web pages

In this module, relevant queries are captured from the search logs to consider queries that are likely to induce users to click frequently on the same set of URLs. The url's are ranked according to most frequently user clicked ones. By providing ranking, we can improve the web pages according to the user needs.

Predicting user behavior

Each query group contains closely related and relevant queries and clicks. By using this approach makes use of search logs in order to determine the relevance between query groups more effectively. In fact, the search history of a large number of users contains signals regarding query relevance, such as which queries tend to be issued closely together, and which queries tend to lead to clicks on similar URLs (query clicks).

IV. IMPLEMENTATION

A seed URLs selection approach is proposed based on user-interest ontology. We can make full use of the user-interest ontology for extracting the user interest topic area and expanding user queries to receive the most related pages as seed URLs.

The output screenshots are shown in Figure 1 & 2 respectively.

Ratings Page

URL	Ratings	Ratings
http:\www.river.com	16	0.42105263
http:\\www.river.com	0	0.0
http:\\www.river.com	0	0.0
http://www.rivernetwork.org/	22	0.57894737
http://www.bcci.org/	0	0.0
http://www.hockey.com/	0	0.0
http:\	0	0.0

Figure 1 : Ratings

The above figure shows the ratings of several web sites.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

User Name	Query	Datetime	SessionCount	ID
selva	river	123	1	1
user1	train ticket booking	08/09/2013 12:11:47	1	2
user1	ticket pnr status	08/09/2013 12:11:47	1	3
user1	cricket score	08/09/2013 12:11:47	1	4
user1	live streaming	08/09/2013 12:11:47	1	5
user1	ticket to coimbatore	08/09/2013 12:11:47	1	6
user1	coimbatore ticket	08/09/2013 12:11:47	1	7

Figure 2 : Clustering

The above figure portrays the clustering of data

V. CONCLUSION AND FUTURE WORK

The seed URLs selection for the focused web crawler is an important research in search engine. In the search engine, it collects an amount of history behaviors which the user visited the search result and forms user log profile. Every record in user log profile contains user query topics and the corresponding clicked URLs. The user-interest ontology construction is proposed by using user log profile. Based on user-interest ontology, we proposed the seed URLs selection approach. In future it may be developed in other domains.

REFERENCES

1. Rungsawang, N. Angkawattanawit, Learnable Crawling: An Efficient Approach to Topic-specific Web Resource Discovery , Journal of Network and Computer Applications , vol. 9, no. 4, pp. 287-296, 2010.
2. D. Vallet, P. Castells, et al., Personalized content retrieval in context using ontological knowledge? IEEE Transactions on Circuits and Systems for Video Technology , Vol. 12, issue no 2, February 2013.
3. A.M. Khattak, K. Latif, S.Y. Lee, Change management in evolving web ontologies, Knowledge-Based Systems vol. 58, issue no. 6, pp. 3041–3052, Jul. 2009.
4. Patel, N. Schmidt, Application of structured document parsing to focused web crawling, Computer Standards & Interfaces , vol . 12, issue no 2, February 2013.
5. RC Chen, CT Bau, CJ Yeh, Merging domain ontologies based on the Word Net system and Fuzzy Formal Concept Analysis techniques , vol. 51, issue no. 1, pp. 229–237, Feb. 2004.
6. Guzman-Arenas, A.D. Cuevas, Knowledge accumulation through automatic merging of ontologies, Expert Systems with Applications , vol. 58, issue no. 6, pp. 3041–3052, Jul. 2009.
7. Harth, J. Umbrich, et al., Searching and browsing linked data with SWSE: the semantic web search engine, Web Semantics: Science, Services and Agents on the World Wide Web 9
8. Sanjanaashree, Accelerating Encryption/Decryption Using GPU's for AES Algorithm, International Journal of Scientific & Engineering Research, Volume 4, Issue 2, February-2013.