Ontology-Based Web Query Classification for Research Paper Searching

MyoMyo ThanNaing
University of Technology(Yatanarpon Cyber City)
Mandalay,Myanmar

Abstract- In web search engines, the retrieval of information is quite challenging due to short, ambiguous and noisy queries. This can be resolved by classifying the queries to appropriate categories. Web query classification is to classify a web search query into a set of intended categories. This system proposed the technique for providing search results by classifying web query. Classifying the web queries into target categories, also known as web query classification is important to improve search relevance. We propose Query Classification Algorithm (QCA) for automatic topical classification of web queries based on domain specific ontology. Here, Ontology is a specialization of concepts in domain and relationships that holds between those concepts. Using ontology as a controlled vocabulary in the process of classification, accuracy value can be improved in retrieving information. The proposed system intends to provide better search result pages for users with interests of intended categories.

Keywords – Information retrieval System, Query Classification Algorithm (QCA), Domain term Extraction, Domain Ontology

I. INTRODUCTION

Search engines have become one of the most popular tools for web users to find their desired information. If user searches information, he has an idea of what he wants but user usually cannot formalize the query. As a result, understanding the nature of information need behind the queries issued by Web users have become an important research problem. Classifying web queries into predefined target categories, also known as web query classification, is important to improve search relevance and online advertising. Successfully classification of incoming general user queries to topical categories can bring improvements in both the efficiency and the effectiveness of general web search.

There are several major difficulties which are needed to consider in query classification. First, many queries are short and query terms are noisy. A second difficulty of web query classification is that a user query often has multiple meanings. Web query classification aims to classify user input queries, which are often short and ambiguous, into a set of target categories. Query Classification has many applications including page ranking in Web search, targeted advertisement in response to queries, and personalization.

In this paper, we propose Query Classification Algorithm, denoted as (QCA), classifies user queries into the intended categories for ranking purpose. After the query classification process, input query is labeled with one or more categories sorted according to their scores Domain ontology is used as a controlled vocabulary. The creation of domain ontology is also fundamental to the definition and use of an enterprise architecture framework. The process of classification queries based on the ontology is presented to improve accuracy value for retrieving information. This intends to provide better search result pages for users with interests of intended categories in top list, for digital library system.

The rest of the paper has been organized as follows. Section II presents the some of the existing techniques related to query classification. Section III describes the overview of the proposed system. The ontology model is discussed in Section IV. Section V is explained about matching to target categories. And then our proposed Query Classification Algorithm (QCA) is explained in Section VI. This paper is concluded in Section VII. Finally, limitations are described in section VIII.

II. RELATED WORKS

Classifying texts and queries are both fundamental concerns in information retrieval. The task of web query classification is to classify queries into a given target category. Lovelyn proposed Web Query Classification based on Normalized Web Distance in [3]. In this system, intermediate categories are mapped to the required target categories by using direct mapping and Normalized Web Distance (NWD). The feature set is the set of intermediate categories retrieved from a directory search engine for a given query. The categories are then ranked based on three parameters of the intermediate categories namely, position, frequency and a combination of frequency and position. In [4], the system proposed Taxonomy-Bridging Algorithm to map target category. The target categories typically does not have associated training data, the KDD CUP 2005 is used. The Open Directory Project (ODP) is used to build an ODP-based classifier. This taxonomy is then mapped to the target categories using Taxonomy-Bridging Algorithm. Thus, the post-retrieval query document is first classified into the ODP taxonomy, and the classifications are then mapped into the target categories for web query.

The system is considered to address the problem of query classification by using conditional random field (CRF) models in [1]. This system uses neighboring queries and their corresponding clicked URLs (Web pages) in search sessions as the context information. The system is not able to find a search context if the query is located at the beginning of search session.

Beitzel exploits both labeled and unlabeled training data for web query classification system in [5]. Diemert and Vandelle propose an unsupervised method based on automatically built concept graphs for query categorization in [6].

Ernesto William presents an approach to classify search results by mapping them to semantic classes that are defined by the senses of a query term. The criteria defining each class or 'sense folder' are derived from the concepts of an assigned ontology in [12]. Some work has been dedicated to using very large query logs as a source of unlabeled data to aid in automatic query classification. In our proposed approach, domain ontology is used as controlled vocabulary for query classification. This system combines the query classification algorithm with the benefits of statistical approaches based on IR techniques.

III. OVERVIEW FOR THE PROPOSED SYSTEM

The aim of query classification is to classify a user query Qi into a list of n categories ci1, ci2,…cin, where cij selected from set of N categories {ci1, ci2,….cin} [4]. Among the output ci1 is ranked higher than ci2 and ci2 higher than ci3 and so on. In this system, the process we used to provide categorical information of search query is described in the following;

1. The user types his query.

2. Domain terms of input query are extracted domain terms by using N-grams algorithm.

3. These search terms are input of Query Classification Algorithm (QCA), which is used to label the input query into the user intended categories.

4. Intermediate categories or search result categories are matched to target categories by using domain ontology.

5. After the query classification process, search result documents are ranked according to the scores of categories predicted by a QCA algorithm.

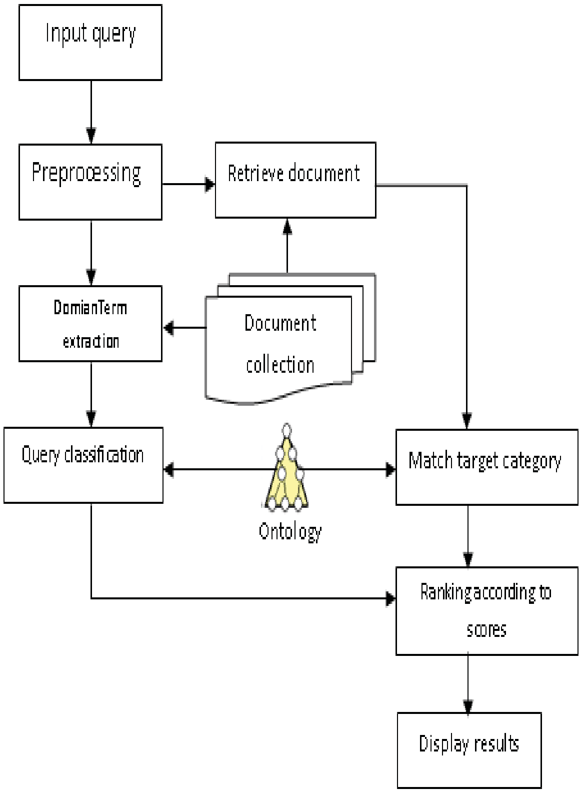


Figure 1. The architecture of proposed system

To get domain terms, at first, user query is preprocessed. Then, domain terms of input query are extracted by using N-Gram term extraction algorithm from Domain corpus. This algorithm works incrementally by first computing the frequency of 1-grams and then considering n-grams of increasing length, each time keeping those which occur with a frequency above a threshold [7]. For this system, trigram is used to get domain terms of input query.

The architecture of proposed system is shown in figure 1. User given query is passed into the search engine. Each result document is belonged to one or more categories. Result categories or intermediate categories are matched to target categories by using domain ontology. The result documents are then ranked according to the scores of important categories predicted by the Query Classification Algorithm instead of term frequency.

IV.DOMAIN ONTOLOGY MODELING

Ontology renders shared vocabulary and taxonomy which models a domain with the definition of objects and/or concepts and their properties and relations. Using ontology as a controlled vocabulary, accuracy value can be improved in retrieving semantic information. In here, ontology is an information model containing concepts and relations in the area of computer science as our case study. We assume ontology is organized as directed acyclic graphs. Each node represents a class and there is relation between them.

In construction of ontology model, concept and property relationship in professional field are defined and field ontology is constructed based on [9] and [10], according to the professional field (Computer Science) as shown in Table I. There are categories in computer science domain encoded as classes such as Artificial Intelligent, Network Technology, Data Mining, Software Engineering, and Information system are examples of some classes. These categories consist of several subcategories or subclasses. For example, Artificial Intelligent has subcategories such as AI Learning, Expert System, Natural Language Processing, Robotics, Deduction and Theorem Proving and so on. Each instance has values. Ontology is applied not only in the process of query classification to get the concepts of each term but also to match target category. Figure 2 shows example of the class, instances and values of domain ontology. The terms from ontology are queried to further process by using SPARQL 1.1 language as in Table II.

TABLE I: Example of class and relationship of Artificial Intelligent

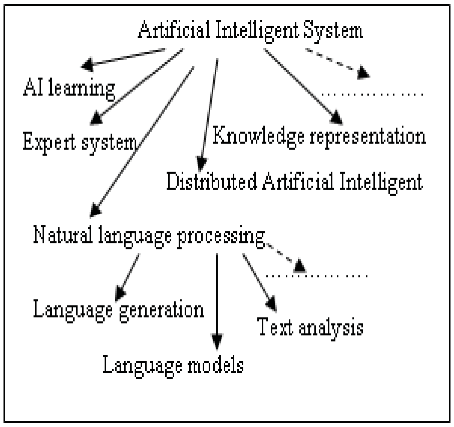
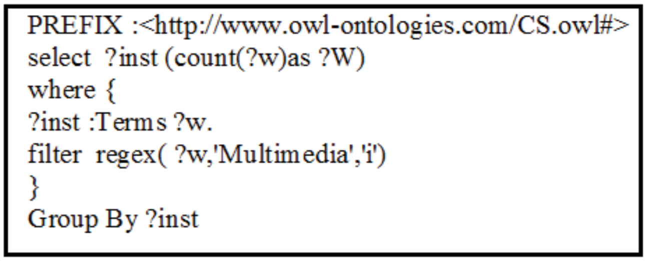
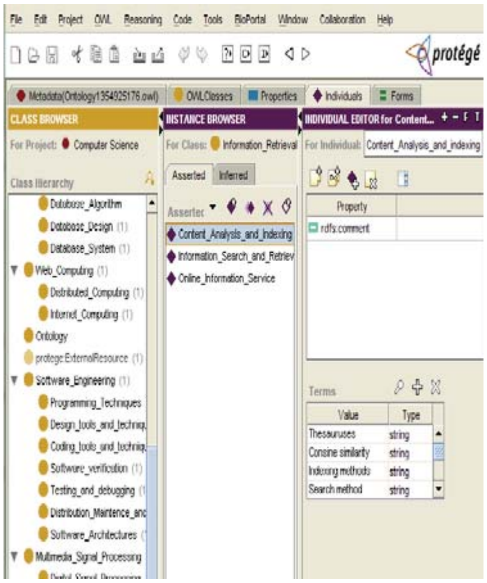


TABLE II: Example of term extraction using SPARQL query language



```
PREFIX :<http://www.owl-ontologies.com/CS.owl#>
select ?inst (count(?w)as ?W)
where {
?inst :Terms ?w.
filter regex( ?w,'Multimedia','i')
}
Group By ?inst
```


File Edit Project OWL Reasoning Code Tools BioPortal Window Collaboration Help
protégé
Metadata(Ontology1354925176.owl)
OWLClasses
Properties
Individuals
Forms
CLASS BROWSER
INSTANCE BROWSER
INDIVIDUAL EDITOR for Content...
Class Hierarchy
Asserted
Inferred
Asserted
Property
rdfs:comment
Terms

For Project: Computer Science

For Class: Information_Retrieval

For Individual: Content_Analysis_and_indexing

Database_Algorithm
Database_Design (1)
Database_System (1)
Web_Computing (1)
Distributed_Computing (1)
Internet_Computing (1)
Ontology
protege:ExternalResource (1)
Software_Engineering (1)
Programming_Techniques
Design_tools_and_techniq
Coding_tools_and_techniq
Software_verification (1)
Testing_and_debugging (1
Distribution_Maintence_and
Software_Architectures (
Multimedia_Signal_Processing

Content_Analysis_and_Indexing
Information_Search_and_Retriev
Online_Information_Service

Value	Type
Thesauruses	string
Consine similarty	string
Indexing methods	string
Search method	string

Figure 2. Example of the class, instances and values of domain ontology

V. MATCHING TO TARGET CATEGORIES

Each document can be tagged with one or more categories. These categories are used as intermediate categories to match target categories. After matching, search results are labeled with target categories. In this system, ontology is used for matching techniques. Here, ontology is organized as directed acyclic graphs. Each node represents a class and there is relation between them. Therefore, ontology provides hypernym/hyponym hierarchies according to the relationship of classes. For example, Expert system, Natural Language Processing, AI Learning is subcategory of Artificial Intelligent and Digital signal processing and Image processing is subcategory of Signal processing. By using hypernym/ hyponym taxonomy of ontology, there are two ways for matching to target categories.

(1) If intermediate categories are in the same category of ontology, then retrieve hypernym of these until it is the parent of corresponding category as the target category. For example "Expert System\ Natural Language Processing\ AI Learning" is matched to Artificial Intelligent and Artificial Intelligent is assumed as target category.

(2) If intermediate categories are in different categories, then retrieve hypernym of each until this is parent of each corresponding category.

VI. QUERY CLASSIFICATION ALGORITHM

The task of web query classification is to classify queries into a set of categories. To classify the user query into the user intended categories, we use the domain ontology. Extracted Domain terms of user query are used as input. The matched terms of each domain terms are the set of terms defined in the domain ontology. At first, matched terms of each domain term are extracted. Each can be contained in one or more category and then compute the probability for matched categories.

Step (1): Extracting Matched Terms for each domain terms

Step (2): Probability for each domain terms

```
Input: Domain Ontology O, Extracted domain terms T
Output: $P = \{p_{11}, p_{12}, \ldots, p_{cw}\}$

 Begin
          N(C, T) = 0;
          for eachword t in T
     {
            for each concept c in O
                  if c.contains(t)
                       N(c, t) ++;
      }
          P(C, T) = 1/N(C, T)
          Return
End
```

After computing the probability for matched categories for terms, the value of each category which contains matched terms is calculated in (1).

Step (3): Compute Value (C): the value of particular category containing matched terms

$$Value(C)= \frac{P(C,T)\times \text{no of matched terms for particular category}}{\text{Total no of matched terms}} \quad (1)$$

For more than one domain terms, the system decides important categories by summation the value of same category for all terms in (2).

Step (4): Compute Score(C): the score of each category for all domain terms.

$$Score(C) =\sum_{C=1}^{n} groupbyCategory(Value(C)) \quad (2)$$

In example, user input query is “Query Process of Natural Language statement Using Metadata” and the steps are shown in below,

User query: “Query Process of Natural Language statement Using Metadata”

Domain Terms: “Query Process”, “Natural Language”, “Metadata”

Step (1): Query processing, Natural language processing, Natural language processing, Natural language, Natural language interfaces, Metadata

Step (2): For the term “Query process” relates to Intelligent Database category and probability is 1. For “Natural Language”, it relates two categories such as Artificial Intelligent and Information system and the probability of each matched category is 0.5. For “Metadata”, it relates to Intelligent Database category and probability is 1.

Step (3): The value for “Query process” is 1.The values of each category for “Natural Language” is (0.5*(2/3) =0.334) and (0.5*(1/3) =0.167), respectively. The value for “Metadata” is 1.

Step (4): Finally, scores for Intelligent Database is 2, Artificial Intelligent is 0.334, and Information system is 0.167.

VII. Conclusion

We explore the idea of using the concepts in ontology to improve search results for research papers of interested category. In this approach, Query Classification Algorithm (QCA) can provide relevant information for user query. The proposed system is intended to improve accuracy value for information retrieval by classifying user input query as important categories. This can be provided interested search result pages of intended category for users. This system can be used for interested area by using specific domain.

VIII. Limitations

The proposed system can only provide to search and retrieve information about digital library of computer science area. At the present, we are still constructing our ontology for computer science area and developing the system. As future work, we will present the effectiveness of our system with experimental result and performance evaluation.

REFERENCES

[1] H. Cao, D. Hao Hu, D.Shen., D. Jiang, , J.T.Sun, E.Chen,Q.Yang, “Context-Aware Query Classification”,(July 19–23, 2009).

[2] F. Arvidsson;A. Flycht-Eriksson,”Ontologies I"(PDF). http://www.ida.liu.se/~janma/SemWeb/Slides/ontologies1.pdf. Retrieved 26 November 2008.

[3] S. Lovely Rose, K.R. Chandran ,” Normalized Web Distance Based Web Query Classification”, Journal of Computer Science 8 (5): 804-808, 2012

[4] D. Shen, J. Sun, Q. Yang, Z. Chen, “Building bridges for Web query classification”. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, Seattle, WA, USA (2006)131-138

[5] S. Beitzel, E. Jensen, O. Frieder, D. Grossman, D. Lewis,A. Chowdhury, and A. Kolcz,”Automatic web query classification using labeled and unlabeled training data”. InProceedings of SIGIR'05, 2005. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, Salvador, Brazil (2005) 581-582.

[6] E. Diemert, G. Vandelle,: “Unsupervised query categorization using automatically-built concept graphs”. In: Proceedings of the 18th international conference on World Wide Web, Madrid, Spain (2009) 461{47010. [10]Yu, J. and Ye, N.: “Automatic Web Query Classification Using Large Unlabeled Web Pages”, Proceedings of Web-Age Information Management IEEE Computer Society Washington, DC, USA2008.

[7] C. Marques and Anges Braud.: “Mining Generalized Chapter n-Grams in Large Corpora”.

[8] Y. Choueka, N. Dershowitz, L.Tal,“Matching with a Hierarchical Ontology”.

[9] http://www.acm.org/class/

[10] http://en.wikipedia.org/wiki/Category:Computer_science

[11] J. Bhogal,A. Macfarlane, P. Smith, “A review of ontology based query expansion”, Information Processing and Management 43,Science Direct,2007, pp :866–886

[12] E.W. De Luca, A. Nürnberger, “Ontology-Based Semantic Online Classification of Documents: Supporting Users in Searching the Web”.