Intelligent Information Retrieval using Query Classification Algorithm

Myomyo Thannaing[1], Ayenandar Hlaing[2]

***Abstract*—**User query classification is an important step for a number of information retrieval. Intelligent Search Service is being developed to help user relevant documents more efficiently unlike traditional Web search engines. We propose Query Classification Algorithm (QCA) for automatic topical classification of web queries based on domain specific ontology. In this system, ontology is an information model containing vocabularies of domain area and relationships that holds between these. In this system, accuracy value is improved by using ontology as a controlled vocabulary in the process of classification. The system aims at contributing to an improved relevance of results retrieved from computer science area without requiring training data set. It is also important to identify the domain terms of search query for classification purpose. Domain term extraction algorithm is used to extract domain terms from user query. The proposed system intends to provide better search result pages for users with interests of intended categories.

***Keywords*—** Intelligent Information retrieval, Query Classification Algorithm (QCA), Domain Term Extraction Algorithm, 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. The ontology model is discussed in Section III. The overview of the proposed system is presented in section IV. And then our proposed Query Classification Algorithm (QCA) is explained in Section V and examples of algorithm are in section VI. Section VII is about evaluation performance. This paper is concluded in Section VIII.

II. Related Works

User query classification is an important step for a number of information retrieval. The task of web query classification is to classify user search query into categories. 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

MyoMyo ThanNaing[1] is with the University of Technology (Yadanarpon Cyber City), near Pyin Oo Lwin, Myanmar (e-mail: myomyothannaing@gmail.com).

Ayenandar Hlaing[2], ThanNaing[1] is with the University of Technology (Yadanarpon Cyber City), near Pyin Oo Lwin, Myanmar (e-mail: anandarhlaingucsy@gmail.com)

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 [10]. 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 proposed system combines the query classification algorithm with the benefits of statistical approaches based on IR techniques.

III. 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 [2]. Using ontology as a controlled vocabulary, accuracy value is improved in retrieving information. In here, ontology is an information model containing vocabularies and relation 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 [8] and [9], according to the professional field (Computer Science) as shown in Fig. 1. 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. Fig. 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.

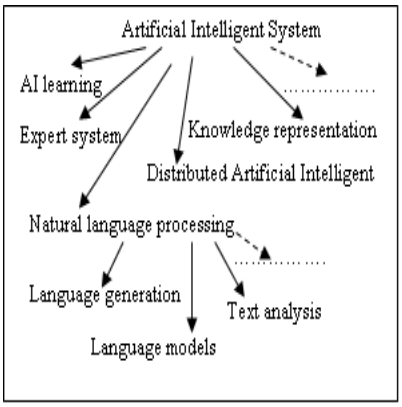


Fig. 1 Example of class and relationship of Artificial Intelligent

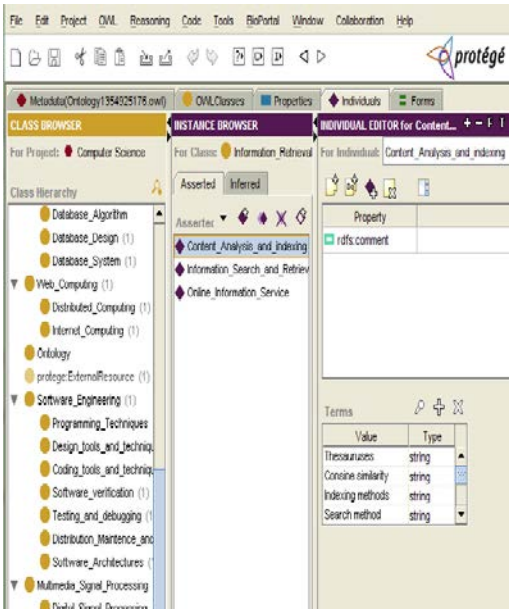


Fig. 2 Example of the class, instances and values of domain ontology

IV. Overview of the Proposed System

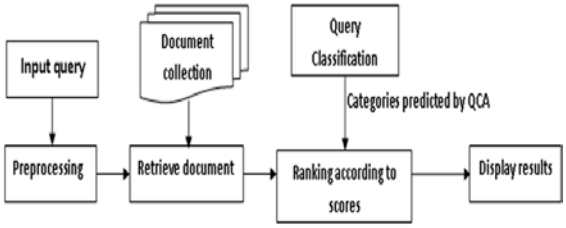


Fig. 3 The architecture of proposed system

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. The architecture of proposed system is shown in Fig. 3. The user input query is passed into the keyword-based search engine which uses TF/IDF approach. Meanwhile, domain terms of input query are extracted by domain term extraction

algorithm. These domain terms are input of Query Classification Algorithm (QCA), which is used to label the input query into the intended categories as described in Fig. 4. After query classification process, the result documents are ranked according to the scores of important categories predicted by the Query Classification Algorithm instead of term frequency.

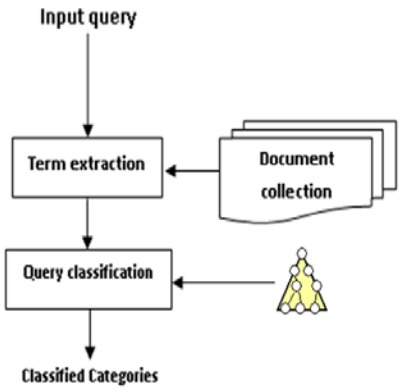


Fig. 4 Query Classification Process

To get domain terms, at first, user query is preprocessed. Domain terms of input query are extracted by using Domain term extraction algorithm from Domain corpus. This algorithm works incrementally by first computing the frequency of 1-grams and then considering 3-grams of increasing length, each time keeping those which occur with a frequency above a threshold. Instead of using fixed size n, by varying the length of gram from of one to three, we get domain terms of all the possible characters appearing in that corpus. It is called the set of generalized character n-grams [7]. For our case study, domain term extraction algorithm is used as shown in Fig. 5. Let C be the domain corpus, *minf* be a threshold, and N(g) be the count of n-gram in C.

```
Input  :    Domain Corpus, query, minf
Output:     the set S of domain terms
Begin
      W1=Tokenize (query)
      W2=StopWords(W1)
      S=W2 with frequency >minf
      i=2;
      Repeat {
            Si= Ø;
            G=PairWords(S,i);
            for each occurance of i-grams g in G
            {
            N(g)=CountofWordsinCorpus(C,g);
            If(N(g)>minf)
            Si=Si+g;
            }
            S=S ∪ Si;
            i=i+1;
      } Until i=3;
      return S;
end
```

Fig. 5 Domain Term Extraction Algorithm

V. Query Classification Algorithm

The task of web query classification is to classify queries into a set of important categories. To classify the user query into the user intended categories, domain ontology is used. 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. Algorithm is explained as bellows.

Input: Ontology, Domain Terms
Output: User intended categories
Begin

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.

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

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

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

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

End

VI. Examples of Algorithm

Here, two examples are shown in below for query classification algorithm. For first example,

User query: "Learning"
Domain term: "Learning"

Step (1): Learning, Concept learning, Parameter Learning, Supervised learning, unsupervised learning.

Step (2): Learning, Concept learning, Parameter Learning relates to category Artificial Intelligent and Supervised learning, Unsupervised learning relate to category Data Mining. So, the term "learning" relates to two categories. The probability of each matched category is defined as 0.5.

Step (3): The values of each category containing matched terms are (0.5 x (3/5) =0.3) and (0.5 x (2/5) =0.2), respectively.

Step (4): Finally, scores for Artificial Intelligent and Data Mining are 0.3 and 0.2.

For second example,
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.

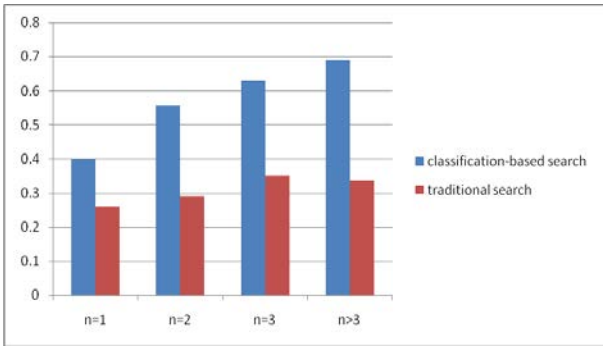
VII. Evaluation Performance

Firstly, we evaluate the comparison of accuracy value for the classification-based retrieval system and traditional search, shown in Table I. For example, in input string "A" that contains "Authenticated and Concealed Data Query in Wireless Sensor Network". In the experiment, there are 2500 total documents. Top 100 of retrieved documents are used to evaluate performance accuracy. In Table II, four set of input queries which contains one, two, three and four domain terms respectively, are tested to evaluate overall accuracy. There are 50 queries in each set. According to the result, if the user query has many domain terms, the proposed-system results more number of relevant documents with user intended category than traditional search. And then, 500 queries test set is used to evaluate classification accuracy. Recall and Precision of the system is 0.899 and 0.893 respectively and F-measure is 89.51%. Correctly classified queries are 447 and incorrectly classified queries are 53.

TABLE I
Comparison of Accuracy Values for The proposed System With Traditional System

Input string	Type of retrieval	No of relevant Documents With categories	Total No of Retrieved Documents	Accuracy Value
A	Traditional search	42	100	0.42
A	Classification -based search	68	100	0.68

TABLE II
Comparison of Overall Accuracy values for the Proposed System With Traditional System



VIII. 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.

Acknowledgment

I would like to thank my supervisor Dr AyeNandar Hlaing, for her help and support through the years. I express my thanks to Dr. Aung Win, Principal of University of Technology (Yatanarpon Cyber City) for granting permission and providing facilities to conduct this thesis. Finally, I would like to thank to all my friends, family, and colleagues for their help and support.

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) .

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

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

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

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

MyoMyo ThanNaing received Bachelor of Computer Science from University of Computer Studies; Yangon in Myanmar .She completed the Master Course from University of Computer Studies since 2008 and especially studied and finished the thesis by Machine learning in Data Mining. Now, she is a Ph.D student in University of Technology (Yadanarpon Cyber City) near Pyin Oo Lwin, Upper Myanmar and mainly study about Software Engineering. Her fields of interest are Ontology and Information Retrieval.