# Context Aware Query Classification Using Dynamic Query Window and Relationship Net

Nazli Goharian
Computer Science Department
Georgetown University, Washington, DC
nazli@ir.cs.georgetown.edu

Saket S. R. Mengle
Dataxu Inc.
Boston, Massachusetts
smengle@dataxu.com

## ABSTRACT

The context of the user queries, preceding a given query, is utilized to improve the effectiveness of query classification. Earlier efforts utilize all preceding queries in the query window to derive such context information. We propose and evaluate an approach (*DQW*) that identifies a set of unambiguous preceding queries in a dynamically determined window to utilize in classifying an ambiguous query. Furthermore, utilizing a relationship-net (*R-net*) that represents relationships among known categories, we improve the classification effectiveness for those ambiguous queries whose predicted category in this relationship-net is related to the category of a query within the window. Our results indicate that the hybrid approach (*DQW+R-net*) statistically significantly improves the Conditional Random Field (*CRF*) query classification approach when static query windowing and hierarchical taxonomy are used (*SQW+Tax*), in terms of precision (10.8%), recall (13.2%), and F1 measure (11.9%).

## Categories and Subject Descriptors

[Pattern Recognition]: Design Methodology- Classifier design and evaluation

## General Terms

Algorithms, Experimentations

## Keywords

Query Classification

## 1. INTRODUCTION

Query classification is the process of assigning predefined categories to queries. Such category information is useful in various domains such as vertical search and advertisement search. The information about the context of the queries is utilized to improve the effectiveness of query classification. We apply two methods, each of which individually or combined (hybrid) increase the effectiveness of existing query classification approaches. Our first method, *Dynamic Query Window* (*DQW*), calculates the query window size based on the unambiguous queries in the query stream. An earlier effort, called *Conditional Random Field* (*CRF*) approach [1], uses a static query window (all preceding queries in the query window) as the context of a given query. We propose an approach that expands the query window dynamically based on the *unambiguous* preceding queries. Unambiguous queries point to only one category. For example, a query that is looking for *virus* is ambiguous (belongs to multiple categories such as *Computer Security* and *Biomedicine*); considering the preceding queries of the query *virus* that are *Flu*, *Fever* and *H1N1*, there is an indication that the user is looking for information on biomedical articles, and subsequently the query is classified as such. In the same query stream, however, the query *H1N1* is unambiguous as it only points to the category *Biomedicine*, and hence, does not require context information for query classification. Preceding queries to query *H1N1* such *as* query *Football, World Cup* and *Soccer*, on other hand, mislead the classification of the query *H1N1*. Thus, our approach utilizes the query stream only for handling the ambiguous queries.

Our second method utilizes the relationships among categories represented in a relationship-net, *R-net*, [4] for query classification. Relationship-net is a network structure where the categories are represented by nodes and relationships among such categories are represented using edges. Relationship-net is automatically generated using text classification [4]. Earlier efforts utilized hierarchical taxonomies to represent the relationships among categories. It was shown in [4] that *R-net* represents more relationships among categories than a hierarchical taxonomy, thus, improving the effectiveness of query classification algorithms. Finally, a hybrid approach that utilizes the *DQW* approach and *R-net* to improve the existing query classification algorithms is presented and evaluated.

Our results indicate that using our approaches, individually or combined, statistically significantly (95% confidence) improve the results of *CRF* approach, which utilizes static query window and hierarchical taxonomy, in terms of precision, recall and F1 measures.

## 2. METHODOLOGY

Our four-step methodology follows.

*Step 1: Determining if context information is needed:* In this step, we identify if a query is ambiguous, so that in the next steps based on the context information of the query, we classify the ambiguous query. We utilize a feature selection algorithm called *Ambiguity Measure (AM)* [3] to assign weight to every query term. $AM(t_i)$ of a term $t_i$, is defined as the maximum *AM* value that a term $t_i$ has in respect to all categories. *AM* of a term with respect to a given category is defined as the ratio of the term frequency of the term in that category to the term frequency of the term in the entire collection. Thus, *AM* assigns higher weight to terms that appear only in one category. The query weight $Wq_j$


*SIGIR'10*, July 19–23, 2010, Geneva, Switzerland.

Figure 1: Evaluation of Precision

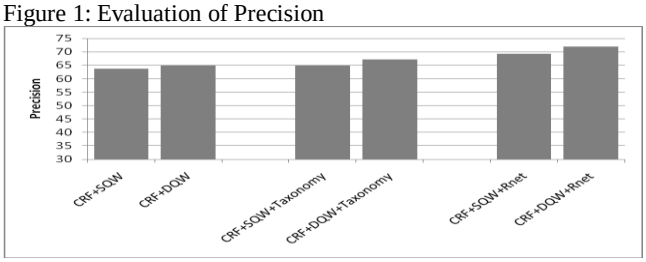



Figure 2: Evaluation of Recall

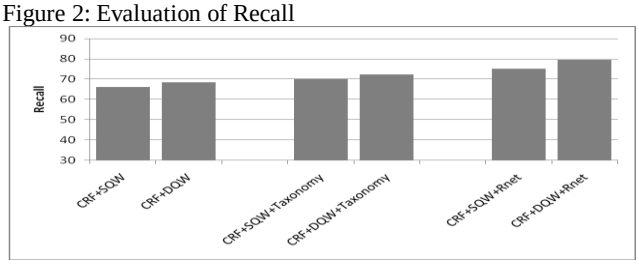
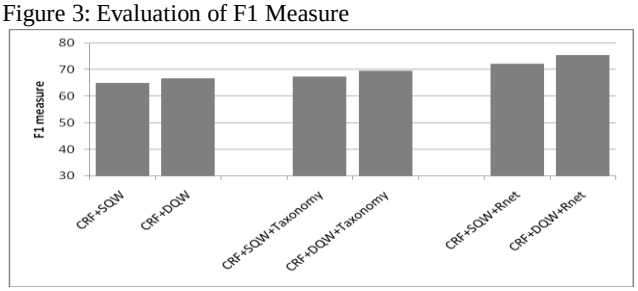



Figure 3: Evaluation of F1 Measure




of query $q_j$ is calculated as the summation of ambiguity measures of all terms $t_i$ in the query (Formula 1), where $T$ is the number of terms in query $q_j$.

$$Wq_j = \sum_{i=0}^{T} AM(t_i) \quad .. 1$$

We only use the context information for queries, whose query weight $Wq_j$ is below an empirically determined threshold (0.7), indicating the query is ambiguous.

*Step 2: Forming Dynamic Query Window (DQW)*: To build context information for an ambiguous query, we need to identify a dynamic query window. Initially, a small static window (three preceding queries) is selected to calculate context information. As unambiguous queries are of our interest in forming context information, we only utilize the information from those preceding queries that their weight ($Wq_j$) is above an empirically determined threshold (0.7). We recursively expand the query window by including three preceding queries of every unambiguous query in the query window. We discontinue expanding the query window when all the three preceding queries of an unambiguous query have weights lower than an empirically determined threshold.

*Step 3: Identifying category relationships using R-net:* In this step, the weights of the queries in the query window are adjusted. The premise is that users search on related categories in the same query stream. Hence, any of the queries in the window whose category is not related to the ambiguous query in hand are penalized by reducing the weights of those queries. Unlike earlier efforts that utilized hierarchical taxonomy *(Tax)* to detect category relationships, we utilize *R-net*, which represents more relationships, to detect relationships between the category of a given query and categories of its preceding queries in the query window. A lower weight ($Wq_j * 0.5$) is assigned to queries in the query window that are not related to the query to be classified. The related queries, however, are assigned the original weight ($Wq_j$).

*Step 4: Classifying ambiguous queries: CRF* is a discriminative probabilistic model used in the classification of sequential data. *CRF* is represented using undirected graph where the vertex represents a random category and edge represents the dependency between two random categories [2]. We utilize *CRF* to classify an ambiguous query into a category based on the categories of the preceding unambiguous queries in a window, based on weights calculated in step 3. We implement the *CRF* approach using an open source utility called CRF++. As the earlier effort [1], using *CRF*, utilizes static query window (*SQW*) and hierarchical taxonomy (*Tax*), we refer to *CRF* approach as *CRF+SQW+Tax*.

## 3. EXPERIMENTAL FRAMEWORK

We utilized 67 categories from KDD Cup 2005 as the set of predefined categories. To build the classifier, we used 500 documents from ODP dataset. The Excite query log is used to extract 500 query streams. Each query stream is at least 5-query long (Query length: avg: 2.7, median: 3). We manually labeled the queries with the KDD Cup 2005 categories. Standard evaluation metrics of precision, recall and F1-measure are used. Statistical significance of our results is measured using paired t-tests.

## 4. Results

Figures 1, 2 and 3 indicate that our hybrid approach (*CRF+DQW+R-net*) statistically significantly (95% confidence) outperforms the existing state of the art (*CRF+SQW+Tax)* approach with respect to precision (10.8%), recall (13.2%) and F1-measure (11.9%). *CRF+DQW* approach, even without using *R-net*, still improves over *CRF+SQW* by 3.5%. Moreover, as *CRF+DQW* expands the window size to allow more unambiguous queries to provide context information, the improvements over CRF+SQW in recall (3%) and F1-measure (3.2%) of query classification are also statistically significant. Using *R-net* along with *CRF* approach (*CRF+R-net*) also statistically significantly improves the precision (6.7%), recall (7.1%) and F1 measure (6.9%) over *CRF+Tax* approach.

## 5. REFERENCES


[1] Cao, H., Hu, D. H., Shen, D., Jiang, D., Sun, J., Chen, E., and Yang, Q., *Context-aware query classification*. SIGIR, 2009

[2] Lafferty, J. D., McCallum, A., and Pereira, F. C., *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* ICML, 2001

[3] Mengle, S., Goharian, N., *Ambiguity measure feature-selection algorithm.* JASIST, 60(5), 2009

[4] Mengle, S., Goharian, N., *Detecting relationships among categories using text classification*. JASIST, 61(5), 2010