

Query Classification using Collective Knowledge



Project Leader(s)

Prof Mark Levene

[mark@dcs.bbk.ac.uk/](mailto:mark@dcs.bbk.ac.uk)

www.dcs.bbk.ac.uk/~mark

Project staff

Zheng Zhu

[Zheng@dcs.bbk.ac.uk/](mailto:Zheng@dcs.bbk.ac.uk)

www.dcs.bbk.ac.uk/~zheng

Other Project Partners

Prof Ingemar J Cox

Project Details

Funded by BBK & UCL

Keywords

Web mining, classification

Project Aims

Understanding the queries submitted to search engines is a key component of modern Information Retrieval. However, due to queries being intrinsically short and noisy, the analysis of these queries is challenging, drawing much attention from current researchers.

In particular, query classification has become an active research field. Query classification benefits not only query routing, thus enhancing vertical search engines, but can also help us to better understand user behaviour and therefore improve the search engine's performance.

Method of Research

We have adopted a two-stage ensemble classifier to construct the query classification system.

In stage 1 we predict the query type as being either navigational, information or transactional. This taxonomy capture the "information need" of the user.

In stage 2 we predict the query class. The class taxonomy is based on a query log file from a search engine. As training data we used a set of manually classified AOL queries from its log. In order to achieve better classification performance, search engine results, search engine related queries, click-through data, social tag information and wikipedia are used to extend the training data.

We employ a Term Match strategy first, which can obtain high precision at the cost of low recall. Then we construct a classifier from several information sources, as shown in

Figure 1. Related concepts and social tag obtained from the collective knowledge are used to enrich the test data, while wikipedia helps us to generate distinct features and feed them into an ensemble classifier. The classifier we have used is bi-gram SVM, which is a state of the art classifier.

The preliminary experiments shows that it can achieve the accuracy of at least 48%, which is quite good for query classification.

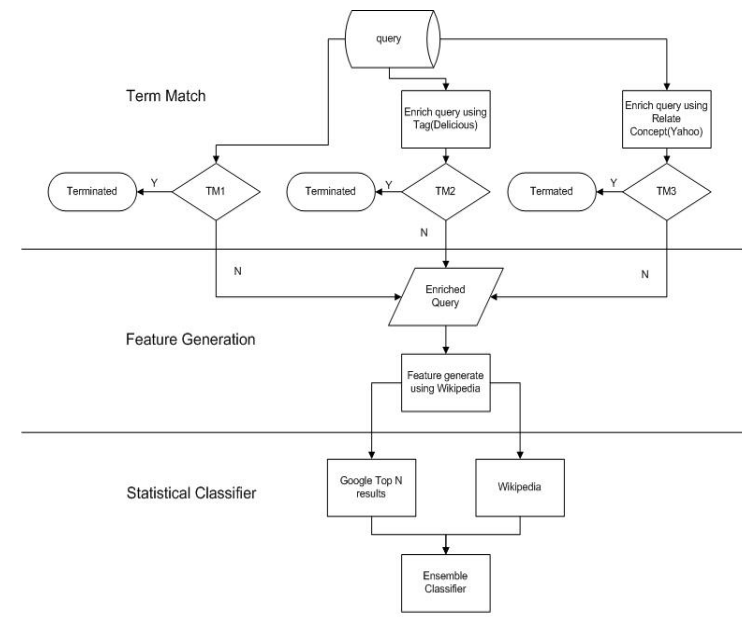


Figure 1. The overview of the Query Classification