ISSN:0975-9646

A Combined Approach Used to find Domain Using Naïve Bayes Classifier

Yogendra Singh Rajput[*1] , Priya Saxena[2]
**1 Research Scholar,CSE,Sanghvi innovative academy,indore ,M.P,India.*
2 Lecturer Scholar,CSE,Sanghvi innovative academy,indore ,M.P,India.

Abstract: This work uses the naive bayes classifier, which is a simple and effective method for establishing classifiers. The proposed model for finding domain, related to user query based on document index matrix. The proposed implementation combine the both approach simultaneously which is term based and phrased based. Document index matrix used term, phrased based document matrix in such a manner that it is compare with training data, and put them into relatively domain. The naïve bayes algorithm used to find maximum probability occurrence from both the matrix. The output comes in the form of suggestion domains list. user easily retrieve the data with minimum time. An experimental result shows the proposed work is better than previous work.

INTRODUCTION

The use of digital text increases as the social media increases their effect in daily life. A number of research groups and individual researchers are working to finding the patterns on these data. This study represents the search engine optimization analysis and sentiment analysis techniques with a new classification algorithm for enhancing the performance of text classification. The given chapter provides an overview of the proposed work and involved investigation.

Knowledge mining is a procedure of mining information from the raw information. Extraction of the similar text from a raw set of text is the generation of text data mining. Clearly, text data belongs to an unstructured method and labelling of information is tricky undertaking, therefore, many of the applications are utilizing the classification approaches for categorizing knowledge.

Text classification captures the relevant result for each user query, Naïve Bayes classifier is a simple and effective approach to classify text document, which uses probabilistic classification technique. Naïve Bayes classifier using Bayes' theorem for classifying unknown retrieved data from google search engine and few modifications are there to increase performance of classifier.

The use of search engines increases rapidly, users of search engine faces problems related to number of replies according to their query. Google search engine returns ranked list of results, which are not relevant to users they are scrolling and finding specific result. The motivation of our study is that queries submitted to a search engine may have multiple meanings. Example, depending on the user, the query "apple" may refer to a fruit, the company Apple Computer or the name of a person, and so forth. Thus, providing categorized query (user, interested in "apple" as a fruit get suggestions about fruit, while users interested in "apple" as a company get suggestions about the company's products) certainly, it helps user's formulate the more effective queries according to their needs.

LITERATURE SURVEY

This section provides the recent efforts and algorithms that are contributing in the small text data processing and accurate sentiment analysis.

Based on implicit observation categorization approaches mainly classified into two fundamental categories: document-based and concept-based. In document-based approach that can be identifying user document preferences from the click through data, to learn a ranking function that optimizes browsing and clicking preferences of a user on the retrieved documents. Joachim's method [16] proposed that a user would retrieve all the search result from top to bottom because of extracting user clicking preferences from the click through data. A ranking SVM algorithm uses the clicking preferences [15] to teach a ranker that best fits preferences of the users. [23] Proposes RSVM that uses the ranking SVM algorithm using a co training framework [14]. Later on, in this paper [2] proposed employing Rank Net to learn clicking and browsing behaviours of users from the click through data.Paper [11] proposed a technique that combined with a novel voting method to determine user's data preferences from the click through data. This technique spying on search data. Concept-based approaches focus on determining the user behaviour.

Paper [1] proposed an algorithm that uses Shannon theory for distribution the entropy to measure the view and visit distribution of domains. When user visit increases that means entropy is increases, while decreasing entropy is a sign of user visits becoming less and suggests the formation of domain preferences.

Paper [18] shows the preferences for independent model for long-term and short-term. The Google Directory used to determine long term preferences, while the short-term preferences are determined from the user's data preferences.

Paper [25] proposed automatically extracting user-interested topics from the user's personal documents that can be browsing histories or emails. The extracted topics are then grouped into a hierarchical user profile (simply called HUP in subsequent discussion), which is to rank the search outcome according to the topical needs of user.

Paper [8] proposed to cluster and organize users' queries into a hierarchical structure of topic classes. A Hierarchical Agglomerative Clustering (HAC) [25] algorithm works in a following ways: Firstly, it employed to construct a binary-tree cluster hierarchy and the second the binary-tree

hierarchy then partitioned in order to create sub hierarchies forming a multidirectional tree cluster hierarchy like the hierarchical organization of Yahoo [6] and DMOZ [3].

Paper [20] proposed an algorithm that combines the group's similar queries according to their semantics. The method creates a vector representation Q for a query q, and the vector Q is composed of terms from the clicked documents of q. Cosine similarity, that applied to the query vectors to discover similar queries.

Paper [21] proposed a method that generates the user profiles automatically based on browsing histories and emails of users' (i.e. personal documents.). The user's interest summarizations into hierarchical structures and from the browsed documents of the users are the frequent terms. This method runs on the assumption that the terms that exist frequently in the browsed documents of the user's represent the users' interested topics. This uses in building hierarchical user profiles that represents the topical interests of the users. Create a user concept-preference profile by considering only the positive preferences of the user. The concepts, extracted for a query in this method. The space covered by these concepts cover more concepts than the actual need of the user. To reflect the users interestingness on the concepts found in the clicked snippets, the weights of the concepts appearing in the clicked snippet are incremented by 1.The other concepts in the concept space which are related to the clicked concepts are incremented based on a similarity score. The concepts that closely related to the concepts clicked (neighbourhood concepts) are incremented to a value close to 1 or 0.The unrelated concepts are assigned weights close to zero.

PROPOSED WORK

The Bayesian Classification represents asupervised learning method as well as a statistical method for classification. Predefined label are there to compare and place it tem comparatively results as well as in statically approach based on hypothesis to bring out the result from the sample based concept in supervised learning method. It can solve diagnostic and predictive problems. This Classification, after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian classification provides practical learningalgorithms and prior knowledgeand observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

Problem Definition

Web data mining is very challenging task to search a desire result what user want. Many filtrations are required to generate or find new domains for user specific query request. Lots of filtration takes lots of time. Many algorithms are solves such problems but every algorithm has its own limitation. No such algorithm proposes optimization problem to solve the problem of search engine with minimum time of unknown dataset. Any algorithm not there to overcome the problem of searching unknown domain.

Proposed Solution

Presented work proposed Naïve Bayes classifier to overcome the problem of search engine optimization of unknown dataset. Proposed solution uses the following steps.

- First, train the entire domain using term wise and phrased wise document matrix.
- At the time of testing, calculate how many terms and phrased used in this web document.
- Apply Naive Bayes classifier to calculate probability of term and phrased of relative document.
- Based on highest probability occurrence compare with training domain and get the result.

Figure 1 shows the proposed model, which describes overall steps in system. First, user gives an input a search query to search engine that returns a list of ranked results then classification applied to that result and user gets categorized results.

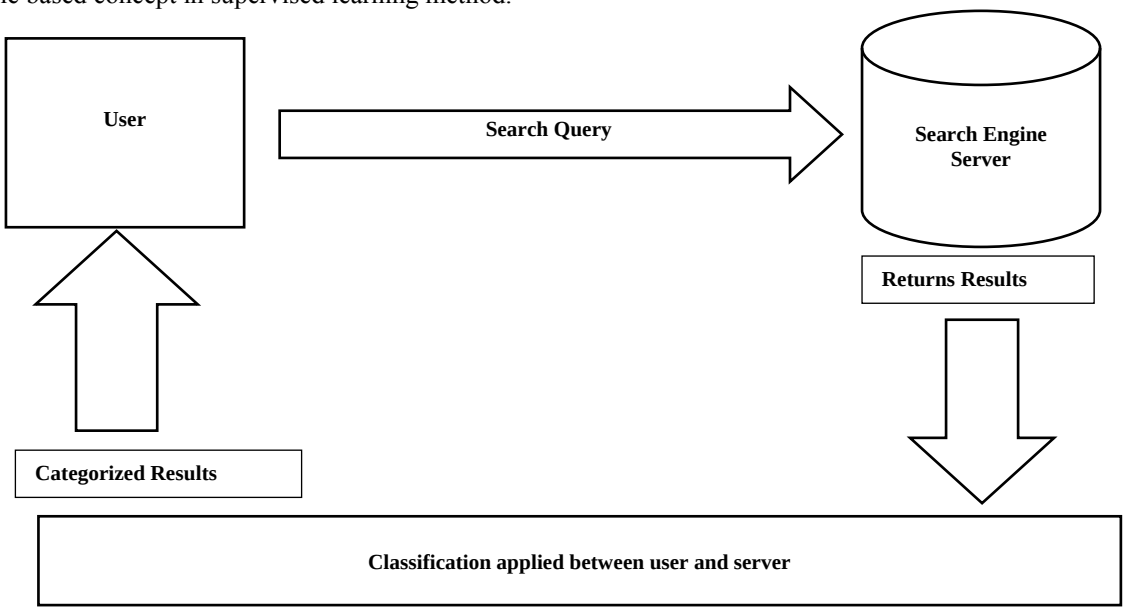

Figure 1 Proposed model

Document matrix stores the record of term and phrased value of different domain in proposed work.

Working Procedure

The main working process of a proposed implementation as follows:

- First, select the categories for training.

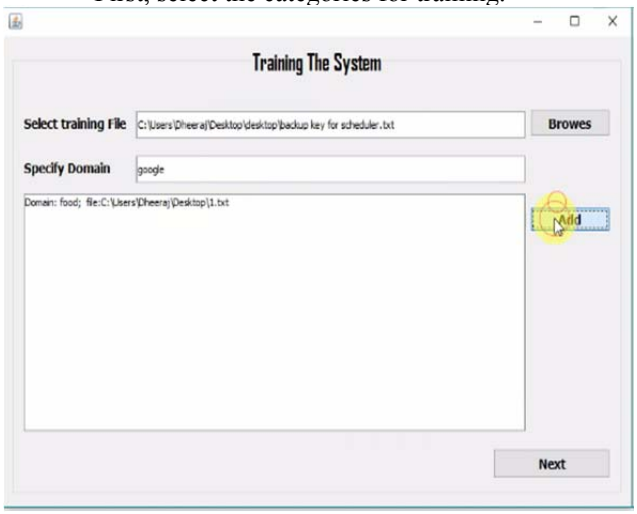

Figure 2 Training System of Proposed Domain Classifier.

This module used to train the system for different domains. Special keyword, which is associated with particular domain, is used for training.

Train the categories using term wise and phrase wise approach, it means single categories will be differentiates by two ways either term wise and phrased wise.

Figure 2 shows the training phase of a proposed classification module.

- After the training procedure, when we give input as a query the related categories found.

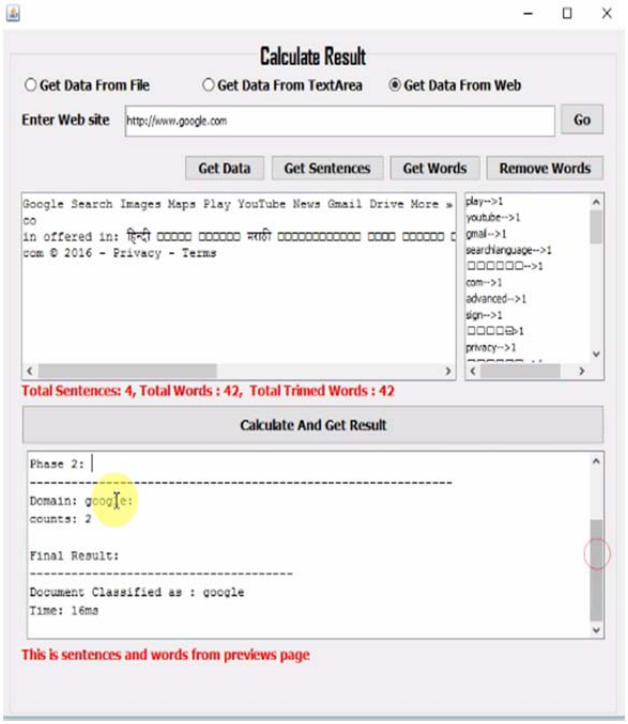

Figure 3 Testing System of Proposed Domain Classifier.

Figure 3 shows the testing procedure of a particular domain finder. It shows the various process involve in this mechanism. When user gives any input document, firstly it will creates the sentences on it and then it will fit into term matrix and phrased matrix.

It will compare based on document matrix and generated the result based on naive bayes classifier.

RESULT ANALYSIS

This section discusses the different types of graph and table related to our result. This shows the comparatively results of simple search and proposed search. The X-axis contains the categories and the Y-axis contains time consumed in terms of milliseconds in these diagrams. According to the comparative results analysis the performance of the proposed technique shows the less time consumption as compared to the traditional technique.

- **Search Result For Java**

Table 1 gives time consumption for query java, the category wise search result related to java query. There are five categories are found related to java keywords which is hardware, internet, multimedia, networks and security. This table gives the various comparisons performed between simple search and using modified Bayes search. The result shows the improved search is better than simple search. Improved search uses minimum time in all aspects.

Table 1 Time consumption for query java

Categories	Simple Search	Improved Search
Hardware	380	270
Internet	370	310
Multimedia	280	230
Networks	210	190
Security	340	260

Figure 4 shows time consumption of the proposed and tradition algorithms for search query java. The graph shows various attributes related to the maximum number of occurrence search result related to java query.

- **Search Result For Package**

Table 2 gives time consumption for query package, which have categories wise search result related to package query. There are five categories are found related to java keywords which is hardware, book, Companies, Auto Racing and Stores. This table gives the various comparisons performed between simple searches and using modified Bayes search. The result shows the improved search is better than simple search. Improved search uses minimum time in all aspects.

Table 2 Time consumption for query package

Categories	Simple Search	Improved Search
Hardware	300	270
Book	310	230
Companies	320	240
Auto Racing	420	210
Stores	230	200

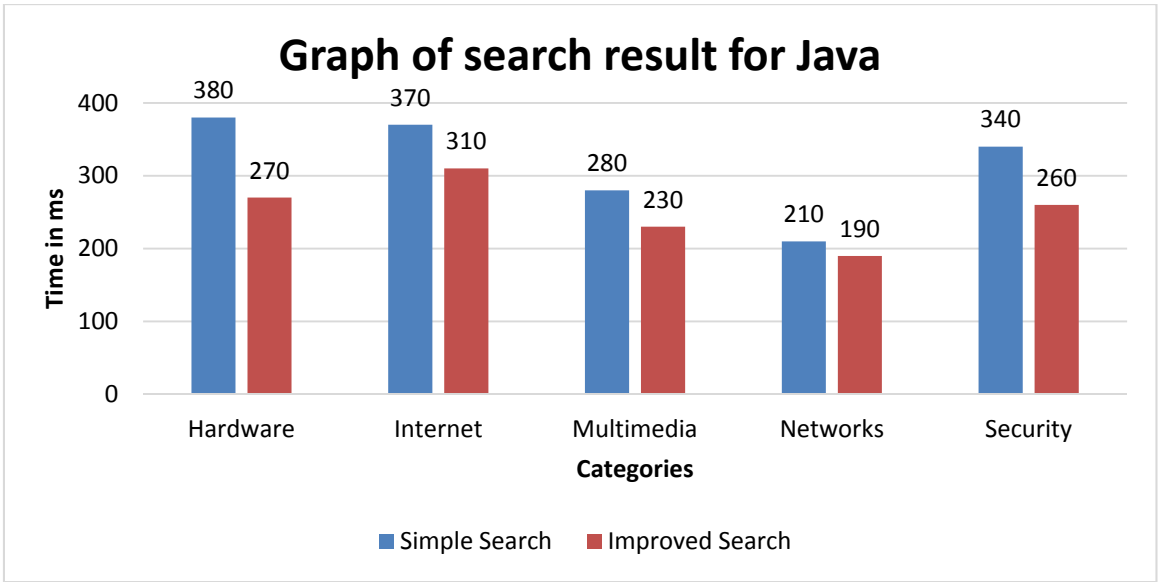

Figure 4 Time consumption of the proposed and tradition algorithms for search query java

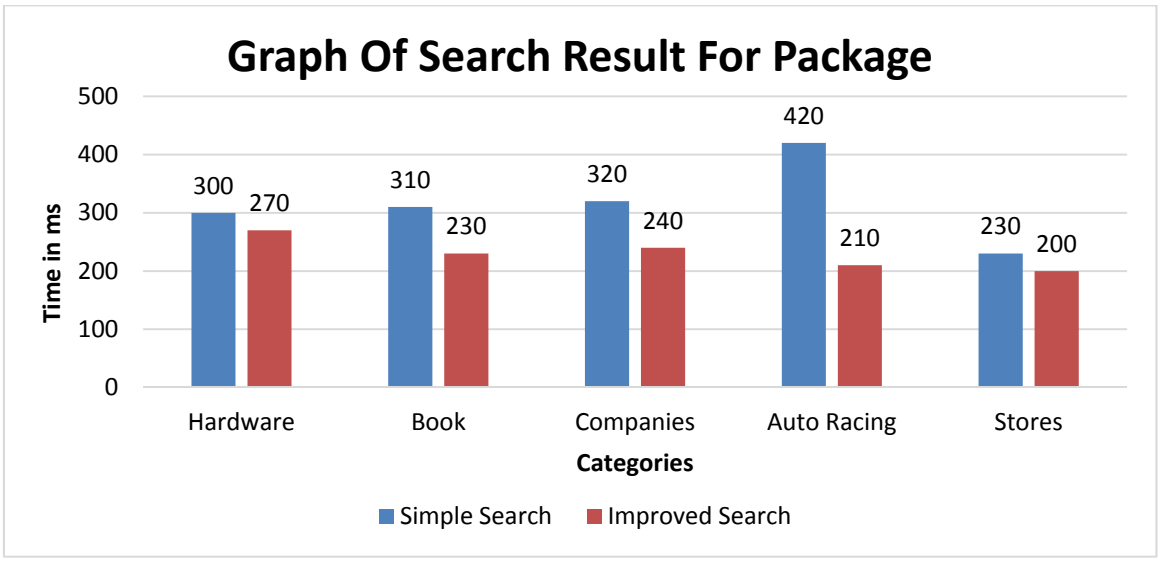

Figure 5 Time consumption of the proposed and tradition algorithms for search query package.

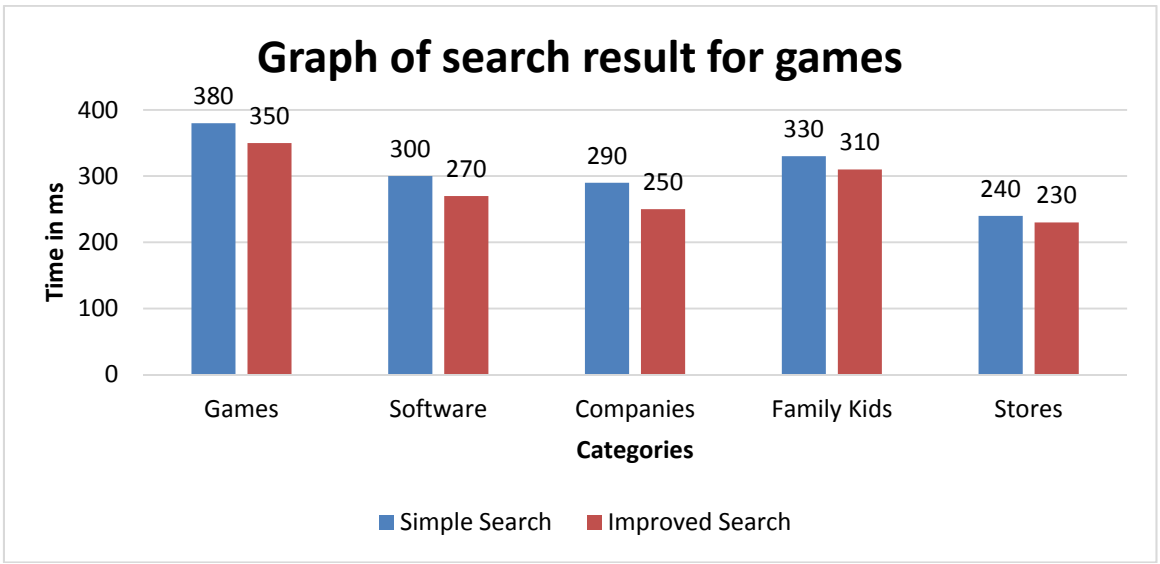

Figure 6 Time consumption of the proposed and tradition algorithms for search query games.

Figure 5 shows time consumption of the proposed and tradition algorithms for search query package. The graph shows various attributes related to the maximum number of occurrence search result related to package query.

- **Search Result For Games**

The table 3 shows the categories wise search result related to Games query. There are five categories are found related to Games keywords which is Games, Software, Companies, Family Kids and Stores. This table gives the various comparisons performed between simple search and using modified Bayes search. The result shows the improved search is better than simple search. Improved search uses minimum time in all aspects.

Table 3 Time consumption for query Games

Categories	Simple Search	Improved Search
Games	380	350
Software	300	270
Companies	290	250
Family Kids	330	310
Stores	240	230

Figure 6 shows the time consumption of the proposed and tradition algorithms for search query games. The graph shows various attributes related to the maximum number of occurrence search result related to Games query.

CONCLUSION

Domain name classification is a great challenge in a web mining. Thousands of domains are there, find the right one from them is quite difficult process. Searching process takes too much time and many filtrations to search out the right one from them. Naive Bayes classifier comes to solve the problem of domain classification. Proposed work uses term and phrased based approach simultaneously to get the accurate result from the training domain. A new query result returns document, when that document enters, Naive Bayes uses to find the probability of highest term and phrased available there using matrix. A modified Naive Bayes algorithm exists to deal with that filter result. Modified Naive Bayes works on selected resulted whose frequency of occurrence is high, so the modified Naive Bayes performance gets higher in terms of timing. Experimental results show the presented work outperforms far better than the previous one.

REFERENCES

[1] Samuel Ieong, Nina Mishra, Eldar Sadikov, Li Zhang, "Domain Bias in Web Search" , ACM New York, NY, USA ©2012.

[2] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, "Ranking Model Adaptation for Domain-Specific Search", IEEE Transactions on Systemsknowledge and data engineering vol. 24, no. 4, April 2012.

[3] Junghoon Chae, Dennis Thom, Yun Jang, Sung Ye Kim, Thomas Ertl, David S. Ebert, "Public behaviour response analysis in disaster events utilizing visual analytics of microblog data",Elsevier Ltd. All rights reserved 2013.

[4] Umajancy. S, Dr. Antony Selvadoss Thanamani, "An Analysis on Text Mining –Text Retrieval and Text Extraction", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013.

[5] Xia Hu, Lei Tang, Jiliang Tang, Huan Liu, "Exploiting Social Relations for Sentiment Analysis in Microblogging", WSDM '13, February 4–8, 2013, Rome, Italy, ACM 978-1-4503-1869-3/02/ 2013.

[6] Rahul A. Patil, Prashant G. Ahire, Pramod. D. Patil, Avinash L. Golande, "A Modified Approach to Construct Decision Tree in Data Mining Classification", International Journal of Engineering and Innovative Technology (IJEIT), Volume 2, Issue 1, July 2012

[7] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD, 1993.

[8] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009

[9] P. Bhargavi, B. Jyothi, S. Jyothi, K. Sekar, "Knowledge Extraction Using Rule Based Decision Tree Approach", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008

[10] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. ACM SIGKDD, 2000.

[11] L. Deng, W. Ng, X. Chai, and D.L. Lee, "Spying Out Accurate User Preferences for Search Engine Adaptation," Advances in Web Mining and Web Usage Analysis, LNCS 3932, pp. 87-103, 2006.

[12] Z. Dou, R. Song, and J.R. Wen, "A Large-Scale Evaluation and Analysis of Personalized Search Strategies," Proc. 16th Int'l World Wide Web Conf. (WWW), 2007.

[13] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, Robert C. Miller, "TwitInfo: Aggregating and Visualizing Microblogs for Event Exploration", Vancouver, BC, Canada. Copyright 2011 ACM 978-1-4503-0267-8/11/2012, CHI 2011, May 7–12, 2011.

[14] M. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real Life Information Retrieval: A Study of User Queries on the Web," Proc. ACM SIGIR Forum, vol. 32, pp. 5-17, 1998.

[15] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. ACM SIGKDD, 2002.

[16] T. Joachims and F. Radlinski, "Search Engines That Learn from Implicit Feedback," Computer, vol. 40, no. 8, pp. 34-40, 2007.

[17] Miloš Radovanović, Mirjana Ivanović, "Text Mining: Approaches And Applications", Abstract Methods and Applications in Computer Science (no. 144017A), Novi Sad, Serbia, Vol. 38, No. 3, 227-234, 2008.

[18] Stefan Stieglitz, Linh Dang-Xuan, "Political Communication and Influence through Microblogging 6 an Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior", 45th Hawaii International Conference on System Sciences, 2012

[19] Leon Derczynski, Diana Maynard, Niraj Aswani and Kalina Bontcheva, "Microblog-Genre Noise and Impact on Semantic Annotation Accuracy", 24th ACM Conference on Hypertext and Social Media 1–3 May 2013, Paris, France Copyright 2013

[20] R.A. Baeza-Yates, C.A. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. EDBT Workshop, vol. 3268, pp. 588-596, 2004.

[21] Andreas Hotho, Andreas Nurnberger, Gerhard Paab, Fraunhofer AiS, "A Brief Survey of Text Mining", Knowledge Discovery Group Sankt Augustin, May 13, 2005

[22] M. Speretta and S. Gauch, "Personalized Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI), 2005.

[23] Q. Tan, X. Chai, W. Ng, and D.L. Lee, "Applying Co-Training to Clickthrough Data for Search Engine Adaptation," Proc. Ninth Int'l Conf. Database Systems for Advanced Applications (DASFAA), 2004.

[24] C.J. Van Rijsbergen, Information Retrieval. Butterworths, 1979. [31] J. Wen, J. Nie, and H. Zhang, "Query Clustering Using User Logs," ACM Trans. Information Systems, vol. 20, no. 1, pp. 59-81, 2002.

[25] Y. Xu, B. Zhang, Z. Chen, and K. Wang, "Privacy-Enhancing Personalized Web Search," Proc. 16th Int'l World Wide Web Conf. (WWW), 2007.