Visualizing Search Results Based on Multi-label Classification

Zhihua Wei Duoqian Miao Rui Zhao Chen Xie Zhifei Zhang
Department of Computer Science and Technology,
Tongji University
Shanghai, China
{Zhihua_wei, dqmiao}@tongji.edu.cn, {zhaorui1, zzf_tj01}@126.com, x_c_grace@163.com

Abstract—**Search engine has played an important role in information society. However, it is not very easy to find interest information from too much returned search results. Web search visualization system aims at helping users to locate interest documents rapidly from a great amount of returned search results. This paper explores visualization of Web search results based on multi-label text classification method. It conducts a multi-label classification process on the results from search engine. In this framework, users could browse interest information according to category label added by our algorithm. A paralleled Naïve Bayes multi-label classification algorithm is proposed for this application. A two-step feature selection algorithm is constructed to reduce the effect on Naïve Bayes classifier resulted from feature correlation and feature redundancy. A prototype system, named TJ-MLWC, is developed, which has the function of browsing search results by one or several categories.**

Keywords-search engine; visualization; multi-label classification; Naïve Bayes; feature selection.

I. INTRODUCTION

As one of the most prevalent applications in today's network computing environment, Web search engines are widely used for information seeking and knowledge elaboration. However, too much returned results make us submerge in the sea of information. The current search engines (such as Google, MSN, Yahoo, etc.) return a sorted search results according to the keywords input by users. In order to find the interesting information, users should browse the title and digest one by one in the search results.

To solve above problem, Web search results visualization systems are designed, aiming at improving the search efficient and accurate and ameliorating users' browsing experiences. Here, Web search results visualization refers to layout search results in a more clear and coherent way. Previous researches mainly concern clustering technology. Related research could be found in literatures [1-4]. The search engines based on clustering technology already exist such as Vivisimo [5] and Grokker [6]. However, there is a problem that the extracted topic of each cluster is far from expressing the main contents of its cluster. In this case, it's difficult for users to locate interesting information in corresponding cluster. This made the process of visualizing search results less significant.

In this paper, we will take into account the diversity of content in returned search results, that is, one entry (a document) may belong to multiple categories. The system adopts a multi-label classifier to arrange the results according to categories. In this case, users could find an interesting document in various classes that it belongs to.

The reminder of this paper is organized as follows. The second section presents previous works. The third section presents our method, i.e. web search results visualization based on multi-label classification and the prototype system, TongJi Multi-Label Web Classifier (TJ-MLWC), by using our method. The fourth section presents the main algorithms in our system: paralleled NB (Naïve Bayes) multi-label classification algorithm and a two-step feature selection algorithm. The fifth section shows the algorithm evaluation and some testing examples of TJ-MLWC. The last section concludes and prospects future works.

II. RELATED WORKS

Categorizing search results is one obvious solution for dealing with information overload. Clustering is one method that allows users to view categorized results without having to deal with the costs and complexities of building taxonomies (see, e.g., the Vivisimo search engine). Zamir and Etzioni made an empirical comparison of standard ranked-list and clustered presentation systems when designing a search engine interface named Grouper, and reported substantial differences in use patterns between the two [4]. Some researchers have experimented with highly metaphorical visualizations. For example, Cugini et al. present users with structural overviews of result sets and promote visualization as the best approach to dealing with broad search tasks [7]. Visualization structures of this type appear to make it easier for users to locate worthwhile information and to comprehend search results.

Some researchers are experimenting with ways of predicting user search intentions, with some testing new ideas on presenting information visually so as to help users locate information more efficiently. Kim and Allen note that cognitive style and task type directly influence search behaviors [8], and Yuan adds that search experiences influence search command decisions [9]. According to Bilal and Kirby, a list of such factors should include user comprehension of the

This paper is sponsored by the National Natural Science Foundation of China (No. 60970061) and Shanghai Leading Academic Discipline Project (No.B004).

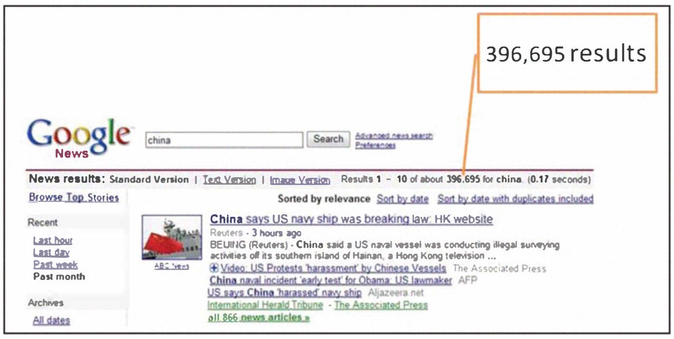
978-1-4244-6789-1/10/$26.00 ©2010 IEEE

search task, individual experience with Web surfing, skill level for manipulating search engines, and the amount of attention an individual gives to a search task [10]. Kao et al. suggest that the concept of thinking style—a distinguishing human factor—should be incorporated into any search engine interface design for better search intention prediction and to help users comprehend search results [11]. Last et al. state that search task type affects users' reactions to hypertext [12].

Newman et al. explored semantic visualizations of Web search results based on topic map [13]. Their topic maps are based on a topic model of the document collection, where the topic model is used to determine the semantic content of each document. The topic model represents documents as a mixture of topics, which is a more flexible representation than k-Means clustering [14], where a document belongs to just one topic. This richer representation means that one can navigate to related articles via any of the topic facets in a given article (this is not possible for k-Means). Furthermore, since the topic model is based on a generative probabilistic model, it can be flexibly extended to include other metadata and attributes, such as authors, citation links, and subject headings, and therefore it is also preferred over Nonnegative Matrix Factorization (NMF) [15].

III. Search Results Visualization Prototype Based on Multi-label Classification

Traditional search engine returns search results one by one ordered by rank value, while there is no class information as shown in Fig. 1. The search engine which we expect is that users could browse interesting contents by clicking corresponding class label, as shown in Fig. 2. Here, the labels in left box are the categories pre-defined. In the right, there are all entries returned by search engine. We could choose one or a few categories, allowing the system to return the entries in specified categories. In this case, the search results are shown in a more clear and coherent approach. Users could save much time in locating the interesting contents and browse them more efficiently.



News results: Standard Version | Text Version | Image Version Results 1 - 10 of about 396,695 for china. (0.17 seconds)

Browse Top Stories

Sorted by relevance Sort by date Sort by date with duplicates included

Recent
Last hour
Last day
Past week
Past month

Archives
All dates

China says US navy ship was breaking law: HK website
Reuters - 3 hours ago
BEIJING (Reuters) - China said a US naval vessel was conducting illegal surveying activities off its southern island of Hainan, a Hong Kong television ...
ABC News
Video: US Protests 'harassment' by Chinese Vessels The Associated Press
China naval incident 'early test' for Obama: US lawmaker AFP
US says China 'harassed' navy ship Aljazeera.net
International Herald Tribune - The Associated Press
all 866 news articles »

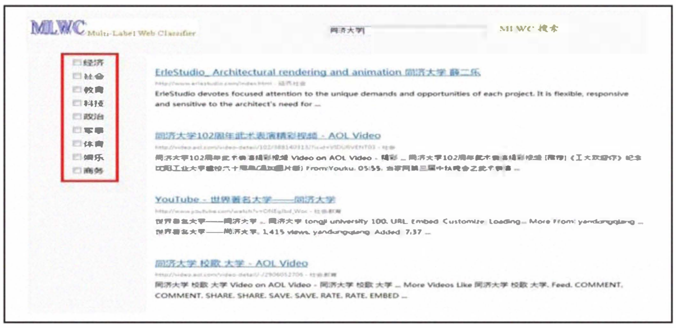
Figure 1. Results of general search engine

According to above consideration, a Web search result visualization system TJ-MLWC based on multi-label classification is constructed. It is composed of three parts: search model, classification model and visualization model.

The search model calls the API of search engine and gets its search results. It also extracts the title and digest of each entry for the using of classification model.

The classification model is composed of classifier learning and classifier testing. Off-line and on-line learning methods are combined to improve the performance of classifier. Firstly, a multi-label NB classifier is trained on a news corpus. Then, the classifier is rectified gradually by users' feedback. This kind of combination ensures the basis classification performance and gradually enhancement of classifier.

The visualization model provides the user interface which is composed of two parts. A Strut 2 is selected to construct the view for users. Apache Geronimo 2.X +Jetty 6 are selected as container. This kind of design ensures satisfying users' requirements, in the same time, reducing the spending of software in arrangement process. AJAX technology is adopted to complete the real-time updating with the changing of class choosing.



□经济
□社会
□教育
□科技
□政治
□军事
□体育
□娱乐
□商务

ErleStudio_Architectural rendering and animation 同济大学 薛二乐
http://www.erlestudio.com/index.html - 经济社会
ErleStudio devotes focused attention to the unique demands and opportunities of each project. It is flexible, responsive and sensitive to the architect's need for ...

同济大学102周年武术表演精彩视频 - AOL Video
同济大学102周年武术表演精彩视频 Video on AOL Video - 精彩 ... 同济大学102周年武术表演精彩视频 [illegible] From:Youku. 05:55. [illegible] ...

YouTube - 世界著名大学——同济大学
世界著名大学——同济大学 ... 同济大学 tongji university 100. URL Embed Customize Loading... More From: yandongqiang ...
世界著名大学——同济大学. 1,415 views. yandongqiang Added 7:37 ...

同济大学 校歌 大学 - AOL Video
同济大学 校歌 大学 Video on AOL Video - 同济大学 校歌 大学 ... More Videos Like 同济大学 校歌 大学. Feed. COMMENT. COMMENT. SHARE. SHARE. SAVE. SAVE. RATE. RATE. EMBED ...

Figure 2. Framework of Web search result visualization

IV. Paralleled Naïve Bayes Multi-label Classification Algorithm

Traditional single-label classification is concerned with learning from a set of examples that are associated with a single label l from a set of disjoint labels L, $|L| > 1$. In multi-label classification, the examples are associated with a set of labels $Y \subseteq L$. Multi-label classification task are generally be decomposed into several single label classification tasks or be solved by transforming some single label classification algorithms. Some algorithms are considered to be effective, e.g., AdaBoost.MH, ML-KNN and so on [16].

In consideration of real-time requirement, paralleled NB classifier is adopted in system TJ-MLWC. NB classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. In simple terms, a NB classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [17]. However, in real classification problems, this assumption is usually hard to be satisfied. Therefore, a two-step feature selection algorithm is proposed in order to eliminate the correlations between features in the process of learning classifier.

For multi-label classification, a set of NB classifiers are learned for each class in parallel. A strategy for predicting the label set of a document is put forward in this section.

A. Two-step feature selection algorithm

Feature selection is a space reduction method which attempts to select the more discriminative features from preprocessed documents in order to improve classification quality and reduce computational complexity. As many words are extracted from documents, we remove stop list words and then perform two-step feature selection algorithm. Firstly, $Text\_freq\_relative_{ij}$ is used to filter the rare features in each class. Then FCBF (Fast Correlation-Based Filter Solution) algorithm proposed [20] is conducted to filter irrelevant and redundant features among classes sharply.

Filtering rare features within class

The number of words in each class is great. However, most of them occur only one or two times. This kind of words is not representative for a class. It is necessary to cancel them before further feature selection. We adopt $Text\_freq\_relative_{ij}$ value to filter them, which is defined as follows.

In corpus D, each text belongs to a class set Y. Here, $Y \subseteq C$, $C = \{c_1, c_2, ..., c_n\}$ is the class set defined before classification. Relative text frequency is noted as $Text\_freq\_relative_{ij}$.

$$Text\_freq\_relative_{ij} = Text\_freq_{ij} / N_i \qquad (1)$$

Here, N_i is the quantity of texts in class c_i in training set. $Text\_freq_{ij}$ is the number of texts which include word j in class c_i. Algorithm 1 is designed to filter rare features within class.

Algorithm 1:

```
For $c_i \in C$, $C = \{c_1, c_2, ...c_i, ..., c_n\}$,
For word $j \in Term_i$,
If ($Text\_freq\_relative_{ij} < \alpha$)
 {remove word $j$;}
Else { $word_j \in Term_i^{'}$;}
$Term^{'} = \{Term_1^{'}, Term_2^{'}, ..., Term_i^{'}, ..., Term_n^{'}\}$
```

Here, $Term_i$ includes all the words extracted in the documents in class c_i, $Term_i^{'}$ includes all the words selected in the class c_i and $Term'$ is the word set in all classes selected by Algorithm 1.

Feature selection among classes

FCBF algorithm proposed a fast feature filter method which could identify relevant features as well as redundancy among relevant features without pair wise correlation analysis [18]. In this algorithm, feature dimensionality is reduced dramatically by introducing a novel concept, predominant correlation based on symmetrical uncertainty [19]. Here, FCBF algorithm is conducted on the feature set $Term'$ gotten in the first step.

B. Paralleled Naïve Bayes Algorithm for multi-label classification

NB algorithm is efficient. For the real-time system, it has a wide application such as in Spam filtering system. For the system TJ-MLWC, classic NB classifier should be adapted to multi-label classification. Classic NB classifier can be described as follows.

For a random document d_j, its feature is $(a_1, a_2, ...a_m)$, here a_i is the i th feature in document d_j. Document class set is $C = \{c_1, c_2, ..., c_k\}$. According to NB classification theory, the conditional probability of document d_j and each class $P(c_i \mid d_j)$ is defined as:

$$P(c_i \mid d_j) = P(c_i)P(d_j \mid c_i) / P(d_j) \qquad (2)$$

Because $P(d_j)$ does not change the result, it can be ignored. $P(d_j \mid c_i)$ could be got from following formula.

$$P(d_j \mid c_i) \approx \prod_{k=1}^{m} P(a_k \mid c_i) \qquad (3)$$

Here, $P(c_i)$ and $P(a_k \mid c_i)$ can be estimated according to following formulas.

$$\hat{P}(C = c_i) = N_i / N \qquad (4)$$

$$\hat{P}(a_k \mid c_i) = (1 + N_{ki}) \Big/ (m + \sum_{k=1}^{m} N_{ki}) \qquad (5)$$

Here, N_i is the amount of texts in category c_i. N_{ki} is the total frequency of word a_k appearing in category c_i.

In order to decide the category of d_j, we could computing the probability of each category c_i according to above formulas when given document d_j. For single label classifier, the maximum of probability of these categories is the predicted category of document d_j.

$$d_j \in c_i \ \ if \ \ P(c_i \mid d_j) = \max_{y=1}^{k} \{P(c_y \mid d_j)\} \qquad (6)$$

We use a parameter P_{thres} to represent the average of posterior probability of document d_j in each class as shown in following formula.

$$P_{thres} = \frac{1}{n} \sum_{i=1}^{n} P(C_i \mid d_j) \qquad (7)$$

When $P(C_i \mid d_j) \geq P_{thres}$, we consider that d_j belongs to class C_i. In this strategy, new document d_j belongs to all classes which satisfy $P(C_i \mid d_j) \geq P_{thres}$. The least amount of labels of a document is "1" when the probability of the document belonging to a class is obviously higher than that of the document belonging to other classes. The maximum amount of labels of a document is "n" when the probabilities of the document belongs to each class are equal.

V. System Testing and Discussion

A. Classifier evaluation

In order to evaluate the effectiveness of the classifier proposed in Section IV, comprehensive experiments have been performed on both English and Chinese corpora. English corpus is coming from NTT communication science research group, which is real Web pages linked from the "yahoo.com" domain [20]. We construct a new Chinese multi-label classification corpus named TJ-Multi-labelCorp1.0 by relabeling part of Sogou Corpus. It consists of 5,800 documents distributed in nine classes such as economy, politics, military, sports, entertainment, science, society, business and education.

Learning process is performed by choosing 70% documents randomly in a corpus. In the first step feature selection, we choose $\alpha = 0.02$ as the threshold based on a large amount of experiments. In this case, about 6,000 features are selected in each class in average. This scale is proper for next step. After the second step feature selection, there are less than 100 features left.

Experiment results both on English and Chinese corpus show that the paralleled NB classifier proposed in Section IV has nearly performance comparing to the famous algorithms (i.e. AdaBoost.MH and ML-KNN). Taking the metric of Hamming Loss [21] as example, the average value of AdaBoost.MH algorithm is 0.0409; the average value of our algorithm is 0.0417. For other frequently used metrics, the case is similar. However, our algorithm has high efficiency, which is very important for real-time applications.

B. System testing

We tested the system TJ-MLWC with the task of displaying Chinese search results by category. System is simulated in a PC (CPU: Intel Pentium 4, Memory: 2G). The time for returning 100 search results is about 5s, which could satisfy browsing requirement.

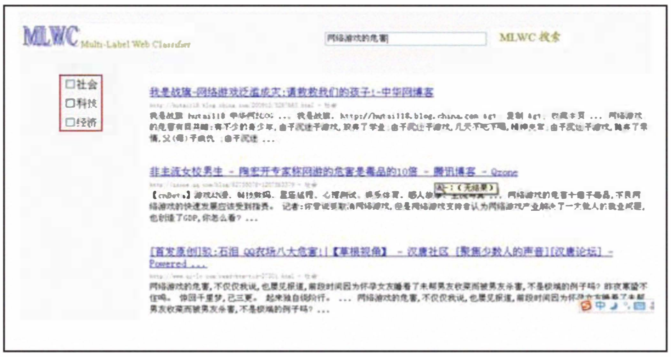
A search example of TJ-MLWC is shown in Fig. 3. The search results of "harm of web game" exist in three classes: society, science and economy. Fig. 4 shows the search results of "harm of web game" when selecting "science" class. They are mainly about the topics on "web psychology" and "web game addiction". These contents are reasonable and acceptable.

C. Discussion

We could have a glance at our search engine TJ-MLWC from Fig. 3 and Fig. 4. From testing process, we could find: (1) TJ-MLWC system could show the search results in one class or several classes by assigning the class label on the left of system interface. (2) If users do not assign the class labels, system return all results like general search engine.

The advantage of this kind of search engine is that user could reduce the browsing range greatly by selecting preferred class labels in left side. In addition, different from the search engine visualization system based on clustering algorithms, a document could appear in more than one class. Theoretically, if a document concerns the contents in several classes, we could find it in all these classes. However, search engine based on clustering is more likely to assign a document in a certain cluster. Usually, the label for a cluster obtained by clustering algorithm is hard to reflect the content of documents in its cluster properly. In this case, the cluster labels are not very effective for user's locating. In TJ-MLWC, class labels are pre-defined to avoid this problem. Moreover, multiple labels ensure that a document could always be found for users having different interests.

There are still some problems to be solved in system TJ-MLWC. Firstly, it mixes the text search results and other kind of search results such as image, video and audio. Further works will construct a hierarchical classifier which classifies files according to their formation first and then conducts multi-label classification on each kind of files. Secondly, the taxonomy is pre-defined in our system. However, with gradually rectification of Web contents, documents in new classes will always appear. How can we rectify the taxonomy automatically according to real demand? All of these problems are our future research objectives.



我是战旗-网络游戏泛滥成灾:请救救我们的孩子!-中华网博客

我是战旗 hutai118 中华网BLOG ... 我是战旗. http://hutai118.blog.china.com &gt; 首页 &gt; [illegible] ... 网络游戏的危害有目共睹:有不少的青少年,由于沉迷于游戏,放弃了学业;由于沉迷于游戏,几天不吃不喝,精神失常;由于沉迷于游戏,抛弃了亲情,父(母)子成仇 ;由于沉迷 ...

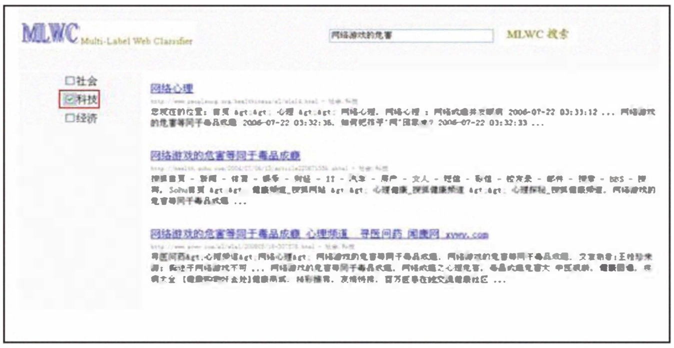
非主流女校男生 - 陶宏开专家称网游的危害是毒品的10倍 - 腾讯博客 - Qzone

【cnDeta】[illegible] ... 网络游戏的危害十倍于毒品,下民网络游戏的快速发展应该受到指责。 记者:你曾说要取消网络游戏,但是网络游戏支持者认为网络游戏产业解决了一大批人的就业问题,也创造了GDP,你怎么看? ...

[首发原创]驳:石祖 QQ农场八大危害!!【草根视角】 - 汉唐社区 [聚焦少数人的声音][汉唐论坛] - Powered ...

网络游戏的危害,不仅仅我说,也屡见报道,前段时间因为怀孕女友睡看了未婚男友收菜而被男友杀害,不是极端的例子吗? 昨夜寒蛩不住鸣。惊回千里梦,已三更。起来独自绕阶行。 ... 网络游戏的危害,不仅仅我说,也屡见报道,前段时间因为怀孕女友[illegible]男友收菜而被男友杀害,不是极端的例子吗? ...

Figure 3. All search results of "harm of web game"



网络心理

您现在的位置: 首页 &gt;&gt; 心理 &gt;&gt; 网络心理. 网络心理 : 网络成瘾并发障碍 2006-07-22 03:33:12 ... 网络游戏的危害等同于毒品成瘾 2006-07-22 03:32:36. 如何把孩子"网"回家中? 2006-07-22 03:32:33 ...

网络游戏的危害等同于毒品成瘾

搜狐首页 - 新闻 - 体育 - 娱乐 - 财经 - IT - 汽车 - 房产 - 女人 - 短信 - 彩信 - 校友录 - 邮件 - 搜索 - BBS - 搜狗. Sohu首页 &gt; &gt; 健康频道_搜狐网站 &gt; &gt; 心理健康_搜狐健康频道 &gt; &gt; 心理探秘_搜狐健康频道. 网络游戏的危害等同于毒品成瘾 ...

网络游戏的危害等同于毒品成瘾 心理频道 寻医问药 闻康网 xywy.com

寻医问药&gt;心理频道&gt;网络心理&gt;: 网络游戏的危害等同于毒品成瘾. 网络游戏的危害等同于毒品成瘾. [illegible] ... 网络游戏的危害等同于毒品成瘾. 网络成瘾之心理危害. [illegible] ...

Figure 4. Search results of "harm of web game" when selecting "science" class

VI. Conclusion

This paper present a new Web search results visualization method based on multi-label classification. A two-step feature selection algorithm and a paralleled NB multi-label classification algorithm are proposed. Experiments both on Chinese and English corpus show that the classifier has nearly performance comparing to some famous multi-label classification algorithm such as AdaBoost.MH and ML-KNN. A prototype system TJ-MLWC is developed. Testing results indicate that this system has the function of browsing search results by category.

The advantage of this kind of search engine is that user could reduce the range of browsing greatly by selecting preferred class labels. It could also avoid the problem of search result visualization system based on clustering methods, i.e., improper cluster labels make it difficult for users to locate preferred contents. Main problem of our system is that taxonomy is hard to extend automatically with new contents appearing. How display different files formation hierarchically is also our future research objective.

References

[1] D. Roussinov, H. Chen. Information navigation on the web by clustering and summarizing query results. Inf. Process. Manage. (IPM) 37(6). 2001. pp: 789-16.

[2] G. Mecca, S. Raunich, A. Pappalardo. A new algorithm for clustering search results. Data Knowl. Eng. 62(3). 2007. pp: 504-522.

[3] N. L. Beebe and Jan Guynes Clark, Digital Forensic Text String Searching: Improving Information Retrieval Effectiveness by Thematically Clustering Search Results, Digital Investigation (4-2) , 2007. pp: 49-54.

[4] O. Zamir, O. Etzioni. Grouper: A Dynamic Clustering Interface to Web Search Results. Computer Networks. vol. 31(11-16),1999. pp: 1361-1374.

[5] Vivisimo Search Engine. http://www.vivisimo.com.

[6] Grokker Search Engine. http://www.grokker.com.

[7] Cugini, J., Laskowski, S., & Sebrechts, M. Design of 3D visualization of search results: Evolution and evaluation. In Proceedings of IST/SPIE's 12th annual international symposium—electronic imaging 2000: Visual data exploration and analysis San Jose, CA. 2000. pp. 23–28.

[8] Kim, K. S., & Allen, B.. Cognitive and task influences on Web searching behavior. Journal of the American Society for Information Science, 52(2), 2002. pp:109–119.

[9] Yuan, W.. End-user searching behavior in information retrieval: A longitudinal study. Journal of the American Society for Information Science, 48(3), 1997. pp:227–229.

[10] Bilal, D., & Kirby, J.. Differences and similarities in information seeking: Children and adults as Web users. Information Processing and Management, 38, 2002 pp: 649–670.

[11] Gloria Yi-Ming Kao, Pei-Lan Lei, Chuen-Tsai Sun. Thinking style impacts on Web search strategies. Computers in Human Behavior archive.Volume 24 , Issue 4 (July 2008). pp: 1330-1341.

[12] Last, D. A., O' Donnell, A. M., & Kelly, A. E.. The effects of prior knowledge and goal strength on the use of hypertext. Journal of Educational Multimedia and Hypermedia, 10(1), 2001. pp:3-25.

[13] D. Newman, et al., Visualizing search results and document collections using topic maps, Web Semantics: Sci. Serv. Agents World Wide Web (2010), doi:10.1016/j.websem.2010.03.005

[14] I.S. Dhillon, D.S. Modha, Concept decompositions for large sparse text data using clustering, Machine Learning 42 (1/2),2001. pp: 143–175.

[15] T.L. Griffiths, M. Steyvers, J.B.T. Tenenbaum, Topics in semantic representation, Psychological Review 114 (2). 2007. pp: 211–244.

[16] G. Tsoumakas, I. Katakis. Multi-Label Classification: An Overview, International Journal of Data Warehousing and Mining, 3(3). 2007. pp:1-13.

[17] D. D. Lewis. Naive (Bayes) at forty: the independence assumption in information retrieval. In: The 10th European Conference on Machine Learning, New York: Springer. 1998. pp:4-15.

[18] Lei Yu, Huan Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003. pp: 856-863.

[19] Press, W. H., Flannery, B. P., Teukolsky, S. A., &Vetterling, W. T. Numerical recipes in C.Cambridge University Press, Cambridge. 1988..

[20] Data set available at http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar.gz.

[21] R.E. Schapire, Y. Singer, Boostexter: a boosting-based system for text categorization, Mach. Learn. 39 (2/3). 2000. pp: 135–168.