http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (8): 1007-1015, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.1007.1015

Support Vector Machine Based Classification of Clicked Document Using Topic Ontology for Profile Generation

[1]S. Prabaharan and [2]R.S.D. Wahidabanu
[1]Anna University, Chennai, India
[2]Govt. College of Engineering, Salem, India

Abstract: User profiles are repetitively described by keyword or concepts space vectors. Unfortunately, profiles are insufficiently or partially interpreted. Hence, search engines are needed to generate a better user profile for providing better search results. This study proposes classifying user queries and clicked documents into topic ontology using machine learning algorithm Support Vector Machine (SVM) for generating a better user profile. SVM divides the document into categories based on the co-occurrences of topics in topic ontology. This method derives the concept based user profile more independently and therefore, it is possible to improve the search engine processes more efficiently. Though, hierarchical categorization together with feature selection has been explained to get improved categorization outcomes. Experimental results show that user profiles are generated once the clicked documents are classified. Finally, the machine learning algorithm SVM is compared with existing query clustering algorithm in order to prove the classification accuracy.

Key words: User profile, SVM, classification, topic ontology, personalization

INTRODUCTION

When queries are issued to the search engine by different users they return the same results irrespective of user interest. The search engine will not be able to provide users' precise needs but it provides in general. Personalization of search engine is not effective but search engine respond with the list of ranked pages based on relevance of the query. So that search engines generate user profiles to identify and get the users' actual needs. Based on the clicked page contents, topic categorization is achieved (Baykan *et al*., 2009). There are several qualities to perform the categorization task with SVM. It illustrates a group of features called a vector. Documents can be classified into topic ontology using SVM algorithm. Categorization of a new clicked document can also be categorized using SVM (Grobelnik and Mladenic, 2005) which contains document as feature vector by changing content of the document into the container of words illustration. Topic ontology concentrates on circumstances when instances are located in any topic ontology node.

At the base of the ontology all documents were situated, documents in leaves characterize the most exact topics. Selecting the most appropriate illustration is called as feature selection. The classification method contains a collection of vectors, each vector denotes a particular node and also it determines weights of the topics and keywords (Grobelnik and Mladenic, 2005). This study introduces clicked document classification and depicts the most important task of classification of clicked documents. The most popular machine learning algorithm SVM is used to assign previously unseen documents to a predefined set of categories and also to classify documents to find interesting information on the web. Clicked documents can be of two categories, relevant or non relevant. In hierarchical classification, document classification has two vital things. They are documents' hyper textual environment and hierarchical formation of the classification set. In concept-based methods, obtain users' topical interests by exploring the contents of the users' browsed documents (Daoud *et al*., 2007) and search histories without human interference (Sieg *et al*., 2007).

The term vector is commonly applied to the concepts and each document is represented as vector. Ontology provides a rich conceptualization of the domain inflects to classification and instances contained by a knowledge base (Middleton *et al*., 2004). Set of keywords and weights could be applied to user's search categories. If a user is interested in searching particular topics (Pretschner and Gauch, 1999), the search is pointed down by rendering recommended outcomes based on user's chosen topics. The concept which is often browsed by the users is extracted from their search record, concept based user profiles and precise users' interests are

Corresponding Author: S. Prabaharan, Research Scholar, Anna University, Chennai, India

mapped into hierarchical structures using topic ontology. Web snippets used to create exact and up to date user profiles. In fact, ontology provides a highly meaningful structure for relating user interests and a rich conceptual among them and allows new topics of interests into the structure (Sieg *et al*., 2007; Pretschner and Gauch, 1999). User profiles hold a set of topics and interests (Shirazi *et al*., 2009).

Ontology is a collection of concepts used to configure information and derive from hierarchical documents. Concept hierarchy is to extend the group of queries and it can be assigned a term weight (Gauch *et al*., 2003). When a user is reading a particular page, they can do some mouse operations. The access time, browsed pages and mouse activities determine a user's interests and content may contain concepts of user's interest. Based on time and concepts the ontology for concept based user profiles are populated (Bhowmick *et al*., 2010).

Ontology is often used to explain the characteristics of profiles. They are commonly proposed for solving heterogeneous users' interests (Vilches-Blazquez *et al*., 2009). Support Vector Machines were introduced to deal with binary classification problems and SVMs are widespread to treat with multi class (Bayoudh *et al*., 2008) problems by concentrating on some binary problems. The main goal is to get personalized user profiles hierarchically in the topic ontology using SVM. In topic ontology, relationship between the topics can be identified and efficient user profiles can be obtained. This paper focuses on classifying user clicked documents for providing a well structured profile for users' search query. The main advantage of topic ontology is that new topics can be easily identified and extracted from documents and added to the topics. This work explains the use of hierarchical design for ordering various group of content. A classification technique typically concerns separating data into two instances. Support Vector Machine classifier is efficient and effective for classification.

Problem assertion: Nowadays because of unexpected growth of the Internet and online accessible documents, the task of classifying documents becomes one of the major problems. Without satisfying the interests of individual users, most search engines are performed in a common way. All of the user profiling methodologies are query based that means a profile is produced for each of the user's queries in previously developed system. But it does not identify the relationship among user's queries and documents. On the contrary, a main approach is proposed for classifying documents into topic ontology, the task which tries to give documents into their particular categories without user interference. Classification technique SVMs were originally proposed to classify the clicked documents into topic ontology, Clicked document classification as a Supervised Machine Learning Problem and Changing documents into Vector Space. The first problem deals with classifying documents into topic ontology. The second problem is about the supervised machine learning. Finally, the third problem is dealing with relationship that exists in the documents and transforming documents into a vector space.

The main contributions of this paper are:

- One of the main intentions is to build precise and complete user profiles to illustrate the user requirements and browsing objectives. Based on users clicking activities the classification is performed and profiles are generated. User profiles are constructed from built in topic ontology based on classification algorithm
- Topic ontology is employed to resolve a semantic problem in which topic ontology catches the semantic information amongst different keywords or topics of documents
- Building a training data set in order to use machine learning algorithm. For extraction of topic related semantic informative documents, Support vector machine is used. Documents are classified into an appropriate class or category by SVM. Isolating top keywords from the complete feature space to decrease the amount of features
- Classifying the document based on relevance and present to the users. Conducting the experiments to evaluate the proposed method and comparing with a query clustering algorithm

User profile can be generated automatically while user is browsing. Using reference ontology, concept weight is applied and user interests are represented. Only by exploring the browsing activities of the user, profiles are created. User visited pages are classified automatically into concepts controlled in reference ontology and consequences of categorization are gathered. Time, length and visited pages are considered in reference ontology to get weights based on the relevant data that the user has surfed (Pretschner and Gauch, 1999). User profiling approaches can be generally classified using query clustering and support vector machine approaches. Using query clustering algorithm, ambiguous queries are classified into diverse query clusters. User profiles are engaged in the query clustering to attain personalization effect. Query concept graph is created by the query

clustering algorithm. User queries and extracted topics are represented as nodes in a graph. User queries are treated as an individual node in the graph. In contrast, SVM method aims at classifying users' documents. Query clustering method treats document as feature vectors and users' documents and surfing activities are planned into a group of topical categories.

User profiles are produced based on classifying clicked documents on the topical categories. Support vector machines are mainly used for classification. It can be applied from text classification to genomic data and also difficult data types than feature vectors (Nyberg *et al*., 2010). Web content mining method using ontology is used for document categorization. This includes formation of ontology for the particular domain, gathering topics, keywords and sub topics of tagged documents, constructing a classification replica, categorization algorithm and categorization of new documents. In categorization algorithm, input phrase extractor with ontology translator is used to make a database from key documents and it can be used as a training data. By having a directory of domain precise stipulations and ontology information, expressions are extracted. To create a group of training documents, user can start the system process by presenting an assembly of domain associated keywords to a search engine (Litvak *et al*., 2007).

Phyu (2009) proposed k-Nearest Neighbor (k-NN) classifiers that are based on learning datasets by similarity. Using n dimensional attribute, training sets are explained and also it specifies a point in an n dimensional gap. If an unknown set is given, a k-NN classifier looks for the place for the k training sets that are neighboring to unidentified group. Neighboring is described as Euclidean distance and unknown set is allocated a most widespread group between its k nearest adjacent. k-NN classifiers are attribute based classifiers. k-NN stores training set and a classifier is not constructed until unlabeled set requires to be classified.

k-NN classifiers allocate equivalent weight to every feature. It possibly will lead to confusion while there are lots of unrelated features in data. As a result, k-NN classifier is used to return real valued calculation of unknown training set. k-NN is responsive to thin organization of data. From a group of substances, the neighbors are taken, however no clear training step is necessary. Substances are symbolized by point vectors in a multidimensional feature space to recognize neighbors (Phyu, 2009; Li and Jain, 1998; Wu *et al*., 2008).

Personalized user profiles: The web search engines make available millions of queries for users while looking for a large set of topics. Each user might have a unique environment and intention for searching particular information through keyword queries. Bedi and Chawla (2010) proposed an integrated approach of personalized web search using Agents and Information Scent. Agent based information retrieval system personalizes the web search by clustering the query sessions of users on the web using information scent, information scent is the measure of the sense of value of clicked web page in the query session with respect to the information need of the user.

A user's interesting profile can be obtained by combining the interesting point group with interesting vector group together, which is denoted by a weighted directed graph (Wu *et al*., 2009). The following methodologies are projected and conversed in this paper. Machine learning algorithm SVM (Deng and Peng, 2006) is used for learning how to classify the documents. SVM utilizes the hierarchical topic structure to decay the classification task. Alternatively, it is interesting that in case of clicked document classification, a linear model is significant adequate to reach fine grained results.

The similarity between document and topics is based on the occurrences of topics. Moreover, related topics based on matching similar topics, it gives the top level concept in the hierarchy. This could help to find out the relations between profiles. Rank search outcomes based on the users' topical interests. A topic hierarchy consists of more conceptual topics on upper levels in the hierarchy. The list of all topics is interrelated with the chosen topic. In this paper, user queries and clicked documents are classified and this process consists of individual user's interests, topical categories of user interests and relationship among topics. Supervised machine learning approach keeps the input and output system. Typically the input space is called as a vector space. The output can be a single real number or Boolean. Machine learning algorithm uses features (i.e.) how many times that document contains the same word. Each feature communicates to measurement of the vector space.

To learn from the original form of clicked documents are not appropriate. So click documents are changed to match the input format of learning algorithm. Since most of the learning algorithms utilize the attribute value depiction, this means converting it into a vector space. First of all, documents must be pre-processed. Document filtering and document stemming are used for neglecting insignificant documents and for decreasing the number of different documents. After that the change takes place (Dumais and Chen, 2000). Documents and queries are signified as vectors in an n-dimensional gap in vector

space representation. Dimension n represents a number of keywords or queries. The related documents contain vector representations close to the query vector. Frequency is calculated based on occurrence of users' query terms and it determines the relationship between query and document.

Based documents' content, categories are predefined in document classification. Hence, each document denotes by an n-dimensional vector called a document vector. To classify the related document into one of the categories, vectors similarity is calculated. Classification is an important task in several information organization and retrieval responsibilities. Classification of clicked document is necessary to focused swarming, to support the web directories improvement, to topic correct web link analysis and to study topical formation of document on web. It improves the search quality. High quality document summaries are able to exactly signify a key topic of a web page. Hierarchical clicked document classification is somewhat inadequate. Based on classical divide and get better classification, the exploitation of hierarchical structure for clicked documents classification is recommended. As per the extension of hierarchy category, documents are updated and classified into topical hierarchy. For extended classification hierarchical SVM is much efficient when compare to flat SVM. It can be very effective and can give fulfilling user needs (Qi and Davison, 2009).

Training data sets have been constructed to facilitate machine learning algorithm. Training data sets are preprocessed before performing the SVM classifier for classifying clicked documents (Materna, 2008). Noise filtering, softening and normalization are done in the preprocessing. It represents the compact of the pattern. Clicked documents are extracted and semantic hierarchy is created for classifying the documents. Semantic hierarchy is used to encode the predefine categories. Document is compared with the categories in a classification process. Most related match or similarity of document is allocated to the categories. Category that has the greatest similarity with the document is classified after verifying all the given categories (Peng and Choi, 2005).

METHODS

Classification method contains the following steps. First, constructing topic ontology to produce user profiles from the topics. Second, the document method, relationships and their algorithms are presented. Thirdly, SVM is exploited and compared with query clustering algorithm.

Topic ontology: Topics are normally ordered in a hierarchical design (Maguitman *et al*., 2010). User profile is generated from user's interested topics. Using keywords and their frequencies topic ontology is constructed. Hierarchical structure presents an understanding of the relations (Li and Jain, 1998; Zhou *et al*., 2006). Concept based user profile is generated from search engine logs based on topic ontology. Spreading activation algorithm was used to optimize the relevance of search engine results. Topic ontology was constructed to identify the user interest by assigning activation values and explore the topics similarity of user preferences. This approach improved the quality of the search engine personalization by identifying the user's precise needs (Prabaharan and Wahidabanu, 2012).

Click based document method: One of the user profiling strategies is click based document approach. Click based document method is aimed at catching up of users' clicking activities. From the clicking document users' document preferences are mined. Users' clicked documents and browsing histories are repeatedly mapped into a set of topical categories. Based on this topical categories user profiles are generated. Most click based document method concentrates on evaluating users' clicking activities traced in the users' click through data. Table 1 is an example of click through data includes an identification of user clicked document and a list of ranked search results returned to the user query and also how many times those documents are actually clicked.

Several personalized systems have been proposed that make use of clicked data to catch the users' interest. By exploring the contents of the users' surfed documents

Table 1: Identification of user clicked document

No. of documents	User Query	Ranked results	Clicked Documents	Extracted topical categories
Doc1	Phone	Phone	Cell phone	Mobile
Doc2	-	Phone Stores	-	Offers
Doc3	-	Products guide	Products guide	Catalog
Doc4	-	phones-meanings mobile, move	-	Dependence
-		-	-	-
-		-	-	-
-		-	-	-
Doc n		SIM details	SIM details	Cards

and histories, users' topical interests have been derived (Leung and Lee, 2010). Click through data in search engines can be of query q, ranked results r and then clicked document c. If a user is submitted a query, ranked results for a query are displayed and then clicked on the links ranked Doc1, Doc2, Doc3 and Doc n.

Identify the relationship among the documents: Relationships are usually represented by semantic relations. Document1 associates to Document2 through relationship R (i.e.,) <D1, R, D2>. Relation consists of extracting the relations clearly represented and complementing the missing relations. Instead of explicitly representing the relation of document D to concept Dn through a relationship such as <D, branch_of, Dn>, ontology relationship in a hierarchical relation between D and a concept called Branch of Dn, i.e., <D, ISA, Branch of Dn>. These two relations are semantically equivalent. This is applied to the branch relationships and is Branch of relationship. To hold for transmit of multiple topics compute a Related Score of every document to every ontology's matching topics:

$$\text{Related score}(t,\text{Top})=\frac{\text{Score}(\text{Dt}).|\text{Dt}\cap \text{d}(\text{Top})|}{|\text{Dt}|} \quad (1)$$

Similarity calculation: Finding the document having the closest match or similarity. The semantic similarity between two documents t1 and t2 is determined by the concept terms Ct1 and Ct2 correspondingly (i.e.,) the frequency (f) in t1 and t2 for users query:

$$\text{Sim}(\text{tl},\text{t2})=\frac{\text{n.df}(\text{Ct1}\cap \text{Ct2})}{\text{df}(|\text{Ct1}|)+\text{df}(|\text{Ct2}|)}/\text{Log n} \quad (2)$$

where, n is the number of documents, df (Ct1) is document frequency of concept term t1 and df (Ct1 n Ct2) is sum of topic relatedness score in document frequency of Ct1 and Ct2 relate to given topics (Stamou and Ntoulas, 2009).

Classification of clicked document: After finding all the given clicked documents, the document will be categorized to category that includes a maximum similarity with the other document. Clicked document classification engages training and testing data sets in topic ontology. Occurrences of frequent queries are extracted to create hierarchical user profiles representing users' topical interests.

New documents are classified using learned classifier during the testing or ready stage. SVM is used as a classifier, because it is very fast and effective for document classification problems (Dumais and Chen, 2000).

Support vector machine: SVM is a supervised machine learning technique used for binary classification. SVM is a binary classifier (Platt, 1999) algorithm which denotes documents using the vector space model (Maguitman *et al*., 2010) and deal with each document as a point in a multi dimensional feature gap. SVM turns around the concept of a margin either area of a hyper plane that splits two data classes. The margin is enlarged and thus generating the major possible space between splitting hyper plane and occurrence on either surface of it has been proven to reduce an upper bound on the expected simplification error. Distance from the hyper plane to adjacent of positive and negative instances is called as margin (Platt, 1999). A linear SVM is hyper plane that splits the group of positive instances from group of negative instances with greatest boundary. In linearly divisible case, enlarging of margin can be expressed as an optimization problem:

$$\min\frac{1}{2}\left\|\vec{W}\right\|^2 \text{Subject to } o_i(\vec{w},\vec{t}_i-b)\geq 1,\forall i, \quad (3)$$

where, t_i is an ith topic instance and o_i is an exact outcome of SVM for ith topic instance. Slack variables are introduced if points are not linearly separable, but deal with severely, points that drop on erroneous part of decision edge. In addition to learn non linear problems, kernel methods are used to change the input. In this paper, linear form of the SVM is proposed to give fine classification accuracy and is rapid to learn and use (Phyu, 2009; Wu *et al*., 2008; Pampapathi *et al*., 2005). To learn the weight of SVM, Sequential Minimal Optimization (SMO) algorithm was proposed by Platt (1999) and it stops the large Quadratic Programming (QP) problem. SVM is not only split the topics, keywords and sub topics into two categories, however make sure that the margin of two categories are biggest. The topics or keywords and sub topics classification are thin and dual. Therefore, SMO algorithm is adequately appropriate for huge feature and topics. New topics can be classified after gone through the weights.

Classification of documents: Figure 1 shows the block diagram of query and clicked document classification using SVM. Classification can be done by calculating w t where w is the vector weight and t is a key vector for a latest document. Because of dual depiction, the sum of the weights for features there in an article. SVM generates subsequent probabilities that are directly equivalent to classes (Dumais and Chen, 2000; Leung and Lee, 2010; Zhuo *et al*., 2008).

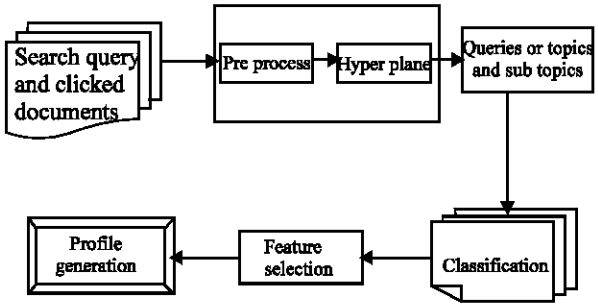

Fig. 1: Classification based profile generation

Ranking: The similarity SIM(t1, t2) is obtained. If it is very much close to topic (Top) and also related significant to documents in topic (Top) the ranking score will be high. Suppose *t* has Related Score (t, Top) and let t1, t2…, tn be the other documents in Top with this t semantically relates with scores of SIM (t, t1), SIM (t, t2), . . . , SIM (t, tn), correspondingly. Where n represents the entire amount of pages in topic (Top) with this t semantically interacts (Leung and Lee, 2010) (i.e.,) SIM > 0:

$$\text{Rank}(t,\text{Top}) = \text{Related score}(t,\text{Top}) + \frac{1}{n}\sum_{i=1}^{n}\text{Sim}(t,t1) \quad (4)$$

Related Score values vary between 0 and 1, with 0 comparable to not at all related and 1 represents category which is greatly meaningful of topic of pages (Stamou and Ntoulas, 2009).

Measurement reduction: Clicked document categorization is frequently described by the numerous numbers of related features in a feature space and rather a little amount of queries or topics and sub topics of each category into topic ontology. Feature dimensionality reduction approach is feature selection. The practice of discovering and eliminating inappropriate and unnecessary features is called as feature subset selection (Li and Jain, 1998).

Feature selection: Feature selection is used when machine learning method is applied for classification. Decrease the feature gap via removing documents that emerge in only a particular document. Numeral techniques for feature selection are compared by Yang and Pedersen, (1997). Calculate the common information between every twosome of feature selections and classifications. The common information CI (FS, C) between a feature selection FS and classification C is described as:

$$CI(FS,C) = \sum_{FS\in(f,\bar{f})}\sum_{C\in(c,\bar{c})} P(FS,C)\log\frac{P(FS,C)}{P(FS)P(C)} \quad (5)$$

The separation of the finest stipulations from the whole feature gap is a method to decrease the features in document classification. In this paper, terms are sorted by weights using individual best features approach. Common information is used as optional to give weights to the terms. In document classification problems, encountering the high dimensional feature spaces are considered and finally, user profiles have been generated successfully.

RESULTS

SVM is a great approach for dealing with machine learning problems. To get an optimal splitting hyper plane, the SVM classifier was used. SVM was utilized as a classifier.

Categories for query clustering and SVM: Mid engine was used for evaluating SVM and query clustering methodologies. Set of 150 topics, sub topics, keywords were gathered and represented by extracting 50 from each of the category in topic ontology. The accuracy of performance measure was evaluated. Based on averaging the accuracy values the results were calculated. Clicked document classification approach was relying on document depiction in a clicked document based input space. Cluster and SVM based representation of topics were presented in terms of classification accuracy.

Table 2 represents the topical categories and their short form for query clustering and SVM algorithms. Most frequent queries were selected and extracted from each class. 20 most frequent queries or topics were chosen for SVM and Query Clustering from each of 7 categories in topic ontology. Some queries occurred in many categories, but only a particular instance for each of these was considered. Total number of all dissimilar queries was 6.

Performance of each of the user's query and users' clicked documents were evaluated. From this, the topics of users' interests were identified. By identifying the users' topical interests user profile was personalized and the performance of search engine was evaluated. The user profiles were employed by the proposed method to group similar queries or topics together according to users' precise needs. The possible motives and differences between the datasets were examined. Based on structure of hierarchical clustering on a clicked document, hierarchy of user interests had been developed. According to constructed topic ontology clicked documents were classified into topic ontology and represented by the vector space model for user profiles.

Table 2: Categories for SVM and query clustering

Category or class	Short form (SVM)	Short form (Query clustering)
Sports	Sport	Spo
Geography and Geo informatics	Geo	Geo
Teachers association	Teacher	Teach
Hospitals and medicines	Hospitals	Hos
Travels and tourism	Travels	Tra
Marine engineering	Marine	Mar
IT professionals	IT	IT

Table 3: Related topics of user clicked topics in documents

Topics	No. of documents	Related topics
Electronics	1000	121
New Inventions	500	23
History Makers	300	12
Sports	750	48
Science	612	26
Cricket	900	53
Total	4062	269

Table 4: Accuracy comparison of SVM, query clustering

Clicked documents	Category or class	Query clustering (%)	Support vector Machine (%)
3D max, Maya, 2D max, Flash, photoshop, multi media	Animation	66.7	75.23
Cricket, Tennis, Foot ball, Hockey, Volley ball, Shuttle cock	Sports	73.1	86.37
C, C++, Java, Dot net, J2EE, COBOL, Mainframe, PHP	Softwares	87	97.62

Table 5: Classification speed and accuracy

	Query clustering	SVM
Speed	Slow	Fast
Accuracy	Better	Best

Performance evaluation: The performance of a classification method was calculated in different ways. An important difficulty with vector support machine approach was that profiling was high measurement of the feature gap replicated by a volume of vectors which characterize the profile. In addition, many features were not only occurring on the ones that incorporate with a particular profile but often it occurred on over all documents. In profile generation so called preprocessing stage was included. Features of all profiles were eradicated to reduce the dimensionality. Users' topics, number of documents that were clicked by the users and related topics were presented in Table 3.

By using the vector space illustration, SVM approach produced higher average classification accuracy correspondingly. The classification precision was also superior when topic symbolized vectors. Different document representations were measured in expressions of classification accurateness. Proposed clicked document classification using SVM with topic ontology was more accurate and usually much faster than query clustering algorithm. Table 4 shows classification accuracy for user clicked documents and their categories.

Comparison of query clustering and SVM: Experiments using the predefined categories and the users' queries or topics were collected from topic ontology for classification using query clustering and SVM. The short explanation and history of these two were provided. Chosen categories from these two were shown. Based on these predefined categories set of topics, keywords and sub topics were collected from each of the web directories. Table 5 shows speed and accuracy for both Query Clustering and SVM.

Using Query Clustering and Support Vector Machine algorithms the classification was tested in order to get an optimal result. 100 topics and 4500 of clicked documents was chosen from topic ontology. Preprocessing methods and machine learning algorithm were compared. Then the greatest resultant method was chosen to test methods. Herein, capacity of SVM approach to achieve a document classification was examined. Query clustering gave 85% of results and accuracy decreased when number of documents increased. The finest results with 97.06% of overall accuracy was obtained with Support Vector Machines algorithm, topic frequency document model and CI-score selection of attributes. Accuracy increased when number of clicked documents increased.

CONCLUSION

This study explains the classification of user clicked document and the general tasks of a clicked document classification into topic ontology. The most popular machine learning algorithm SVM is used for classification. An ontology approach is presented in a way that meets up user necessities. Hierarchical structure of topics easily identified the topics and their relationships. User profile is described as a collection of categories and keywords with weights assigned to the class. Topical categories of user interests in topic ontology and relationship among the concepts are evaluated. This approach provides an accurate concept based user profile which satisfies the users' information needs. SVM has been proven as most powerful learning algorithms for clicked document classification into topic ontology. SVM is rapid and easy for classifying user queries and sub topics into topic ontology.

REFERENCES

Baykan, E., M. Henzinger, L. Marian and I. Weber, 2009. Purely URL based topic classification. Proceedings of the 18th International Conference on World Wide Web, April 20-24, 2009, Madrid, Spain.

Bayoudh, I., N. Bechet and M. Roche, 2008. Blog classification: Adding linguistic knowledge to improve the k-NN algorithm. Proceedings of the 5th IFIP International Conference on Intelligent Information Processing, October 19-22, 2008, Beijing, China.

Bedi, P. and S. Chawla, 2010. Agent based information retrieval system using information scent. J. Artif. Intell., 3: 220-238.

Bhowmick, P.K., S. Sarkar and A. Basu, 2010. Ontology based user modelling for personalized information access. Int. J. Comput. Sci. Appl., 7: 910-911.

Daoud, M., L. Tamine, M. Boughanem and B. Chebaro, 2007. Learning implicit user interests using ontology and search history for personalization. Proceedings of the 2007 International Conference on Web Information Systems Engineering, December 3, 2007, Nancy France, pp: 325-336.

Deng, S. and H. Peng, 2006. Document classification based on support vector machine using a concept vector model. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, December 18-22, 2006, Hong Kong, pp: 473-476.

Dumais, S. and H. Chen, 2000. Hierarchical classification of web content. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, New York, USA., pp: 256-263.

Gauch, S., J. Chaffee and A. Pretschner, 2003. Ontology based personalized search and browsing. J. Web Intell. Agent Syst., 1: 219-234.

Grobelnik, M. and D. Mladenic, 2005. Simple classification into large topic ontology of Web documents. J. Comput. Inform. Technol., 13: 279-285.

Leung, K.W.T. and D.L. Lee, 2010. Deriving Concept-based user profiles from search engine logs. IEEE Trans. Knowledge Data Eng., 22: 969-982.

Li, Y.H. and A.K. Jain, 1998. Classification of text documents. Comput. J., 8: 537-546.

Litvak, M., M. Last and S. Kisilevich, 2007. Classification of web documents using concept extraction from ontologies. Proceedings of the 2nd International Conference on Autonomous Intelligent Systems: Agents and Data Mining, June 3-5, 2007, Russia.

Maguitman, A.G., R.L. Cecchini, C.M. Lorenzetti and F. Menczer, 2010. Using topic ontologies and semantic similarity data to evaluate topical search. Proceedings of the 36th Latin American Informatics Conference, October 18-22, 2010, Asuncion, Paraguay.

Materna, J., 2008. Automatic web page classification. Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2008), December 5-7, 2008, Karlova Studanka, pp: 84-93.

Middleton, S.E., N.R. Shadbolt and D.C. De Roure, 2004. Ontological user profiling in recommender systems. ACM Trans. Inform. Syst., 22: 54-88.

Nyberg, K., T. Raiko, T. Tiinanen and E. Hyvonen, 2010. Document classification utilising ontologies and relations between documents. Proceedings of the 8th Workshop on Mining and Learning with Graphs, July 24-25, 2010, Washington, DC, USA.

Pampapathi, R., B. Mirkin and M. Levene, 2005. A review of the technologies and methods in profiling and profile classification. EPALS Technical Report, 2005. http://www. dcs. bbk. ac. uk/~rajesh/publications/ReviewOfProfilingTechnologies.pdf

Peng, X. and B. Choi, 2005. Document classifications based on word semantic hierarchies. Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, February 14-16, 2005, Innsbruck, Austria.

Phyu, T.N., 2009. Survey of classification techniques in data mining. Proceedings of the International Multi Conference of Engineers and Computer Scientists, March 18-20, 2009, Hong Kong, pp: 727-731.

Platt, J.C., 1999. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In: Advances in Kernel Methods: Support Vector Learning, Scholkopf, B., C.J.C. Burges and A.J. Smola (Eds.). MIT Press, Cambridge, MA, pp: 185-208.

Prabaharan, S. and R.S.D. Wahidabanu, 2012. Ontological approach for effective generation of concept based user profiles to personalize search results. J. Comput. Sci., 8: 205-215.

Pretschner, A. and S. Gauch, 1999. Ontology based personalized search. Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, November 9-11, 1999, Chicago, IL, USA, pp: 391-398.

Qi, X. and B.D. Davison, 2009. Web page classification: Features and algorithms. ACM Comput. Surveys, Vol. 41, 10.1145/1459352.1459357

Shirazi, H.M., M.M. Shirazi and N. Fardroo, 2009. Discovering user interest by ontology-based user profile. Int. J. Intell. Inform. Technol. Appl., 2: 19-24.

Sieg, A., B. Mobasher and R. Burke, 2007. Ontological user profiles for personalized web search. Proceedings of the 5th Workshop on Intelligent Techniques for Web Personalization, July 23, 2007, Vancouver, British Columbia, Canada.

Stamou, S. and A. Ntoulas, 2009. Search personalization through query and page topical analysis. User Model. User-Adapted Interac., 19: 5-33.

Vilches-Blazquez, L.M., J.A. Ramos, F.J. Lopez-Pellicer, O. Corcho and J. Nogueras-Iso, 2009. An approach to comparing different ontologies in the context of hydrographical information. Lecture Notes in Geoinformation and Cartography (LNG&C). Information Fusion and Geographical Information Systems, pp: 193-207, http://iaaa. cps. unizar. es/curriculum/08-Publicaciones-Articulos/art_2009_LNGC_Approach To Comparing. pdf

Wu, X., V. Kumar, J.R. Quinlan, J. Ghosh, D.J. Hand and D. Steinberg *et al*., 2008. Top 10 algorithms in data mining. Knowledge Inform. Syst., 14: 1-37.

Wu, Z.Z., Q.T. Zeng and X.W. Hu, 2009. Mining personalized user profile based on interesting points and interesting vectors. Inform. Technol. J., 8: 830-838.

Yang, Y.M. and J. Pedersen, 1997. A comparative study on feature selection in text categorization. Prceedings of the 14th International Conference on Machine Learning, (ICML'97), Morgan Kaufmann, pp: 412-420.

Zhou, X., S.T. Wu, Y. Li, Y. Xu, R.Y.K. Lau and P.D. Bruza, 2006. Utilizing search intent in topic ontology-based user profile for web mining. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, December 18-22, 2006, Hong Kong, pp: 558-564.

Zhuo, L., J. Zheng, X. Li, F. Wang, B. Ai and J. Qian, 2008. A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. Proceedings of the Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images, June 28-29, 2008, Guangzhou, China.