Automatic Retrieval of User Interests based on Tags in Social Media

[1]D.Sneha, [2]K.F Bharati

[1]M.Tech (CSE), [2]Assistant Professor

Department of CSE, JNTUACE, Anantapur, A.P.

Abstract:

Social media provides an environment of information exchange. They principally rely on their users to create content, to annotate others content and to make on-line relationships. The user activities reflect his opinions, interests, etc. in this environment. We focus on analyzing this social environment to detect user interests which are the key elements for improving adaptation. This choice is motivated by the lack of information in the user profile and the inefficiency of the information issued from methods that analyze the classic user behaviour (e.g. navigation, time spent on web page, etc.). So, having to cope with an incomplete user profile, the user social network can be an important data source to detect user interests. The originality of our approach is based on the proposal of a new technique of interests detection by analyzing the accuracy of the tagging behavior of a user in order to figure out the tags which really reflect the content of the resources. So, these tags are somehow comprehensible and can avoid tags "ambiguity" usually associated to these social annotations. The approach combines the tag, user and resource in a way that guarantees a relevant interests detection. The proposed approach has been tested and evaluated in the delicious social database. For the evaluation, we compare the result issued from our approach using the tagging behaviour of the neighbours (the egocentric network and the communities) with the information yet known for the user (his profile). A comparative evaluation with the classical tag-based method of interests detection shows that the proposed approach is better.

1. Introduction:

The Web is changing at a very fast pace, whether it be the content versatility or it be the technology that explores the web content in meaningful and useful information. The World Wide Web (Web 1.0) which is primarily based on hyperlinks requires keywords, co-occurrence and page rank for searching relevant web pages. The relevance of web pages in this face of the Web is usually computed using hubs and authorities (Kleinberg & Lawrence, 2001) or, keyword term frequency (Salton & Buckley, 1987). However, these techniques and other traditional search algorithms besides being simple and computationally sound, lack in searching semantically relevant web pages (Navigli & Velardi, 2003). This means that the pages that contain synonyms, hypernyms or hyponyms for the keywords rarely get incorporated during the search. The World Wide Web (WWW) being the first version of the Web is usually referred as 'Web' in early literature. However, in the thesis the Web has been used for the existing Web consisting of the WWW, the Social Web and the Semantic Web. To refer a specific version of the Web, these will be referred in particular. As mentioned earlier, the WWW is a vast collection of web pages interconnected with each other through web links called hyperlinks. Web page, a unit of information on the WWW has many synonyms like document, resource and page (Sebesta, 2007). These synonyms have been used interchangeably depending on the context in the thesis. Information retrieval has attained new definitions with the advent of the Web. The web information retrieval deals with the representation, storage, organization of, and access to information items (Baeza-Yates & Ribeiro-Neto, 1999). Low cost, greater access, publishing freedom and linking documents to many other documents on the Web are the primary reasons for the popularity of the Web as a highly interactive medium and immeasurable source of information. Searching useful information to users' interest in the ever-growing volume of the Web is a real challenge for Web information retrieval research.

In a conventional information retrieval system a document is described logically as a collection of index terms. An index term is a keyword which has some meaning of its own such as nouns. In general, the index term may consist of all words in text of the document. However, considering index in this way raises concerns over the text semantics. This issue has been discussed many times in the information retrieval literature. These index terms are compared to find similarity or relevance of the document to a query using various models. The web retrieval process can be explored in one of the two operational modes, ad-hoc and filtering. In ad-hoc retrieval, the documents in the collection remain relatively static while new queries are submitted to the system. In the other mode, the queries relatively remain static while new documents come in the system (and/or leave the system). This operational mode is termed as filtering. The work in the thesis belongs to the filtering task. In filtering, the ranking of documents is based on the users' information need, which is usually constructed through a set of keywords provided either by a user explicitly or extracted implicitly through some preferred relevant documents. This initial information need to improve searching is sometimes referred as 'user profile' or expansion of the query/topic which takes care of a user's information needs. The simplistic way to construct the expanded topic list is to ask user to provide keywords related to his/her search requirement. In some cases the user is also asked to provide relevance feedback about the searched documents to build a training set (consisting of two sets of relevant and non-relevant documents) to be used for improving future retrieval results. Though this approach is simple, it requires a user to provide lot of details that describes his/her profile. Moreover, the user is expected to be familiar with the search topic. She/He has to provide related keywords or required to be able to judge the relevance of documents. The work in the thesis has adopted an approach to construct an expanded topic list for filtering by using semantic knowledge on a search topic. Constructing expanded topic list using semantic structure (consisting of the search topic and its useful related concepts) has a number of benefits towards the filtering task as compared to the above mentioned method. The semantic structure based topic expansion methods alleviates the need for the user to provide related keywords on the required search topic and neither the user is required to spend time in constructing the training set. Context sensitive document retrieval is the added benefit of the semantic structure based filtering.

2. Related Work:

Feature Selection (FS) is a search process in the field of data mining which selects a subset of salient features to build learning paradigm such as decision trees and neural networks. Some irrelevant and/or redundant features usually exist in the training data which makes learning tougher and also degrades the performance of trained model. More precisely, good FS techniques can detect and ignore noisy and false features. This process leads to increasing the quality of dataset after feature selection. Two quality factors need to be considered here: relevancy and redundancy. A feature is said to be relevant if it is prognostic of the decision feature(s); else it is irrelevant. A feature is deemed to be redundant if it's correlation with other features is high. An informative feature must be highly correlated with the decision concept(s), but it is highly uncorrelated with others. Many feature selection algorithms are involved in heuristic or random search methods in order to decrease the time complexity.

In [2], feature adaptation techniques to retrieve more relevant images are present. It is an effective feature space dimension reduction according to user's feedback, but also improves the image description during the retrieval process by introducing new significant features. Feature-Adaptive Relevance-Feedback (FA-RF) uses two iterative techniques to make use of the relevance information that is query refinement and feature re-weighting. For the adaptation of across RF uses the descriptions of both relevant and irrelevant image, as well as their number and proportions. The query image is located near to the boundary of the relevant cluster in the feature space then the system contains few relevant images. Thus the query refinement mechanism is useful to move the query towards the middle of the cluster of relevant images in the feature space. This FA-RF performs very well in terms of capability in identifying most important features and assigning them higher weights compared with classical feature selection

algorithms. Also maintain compact image description. The main drawbacks are less efficient for large databases. There is also need for an efficient feature extraction algorithm. In [3], a new RF framework is used that combines the advantages of using both the Positive Example (PE) and the Negative Example (NE). This method learns image features and then applies the results to define similarity measures that correspond to the user judgment. The use of the NE allows images undesired by the user to be discarded, thereby improving retrieval accuracy. This method tries to learn weights the user assigns to image features and then to apply the results obtained for retrieval purposes. It also reduces retrieval time. It clusters the query data into classes and model missing data, and support queries with multiple PE and/or NE classes. The main function of this method is that it assigns more importance to features with a high likelihood and those which distinguish well between PE classes and NE classes. The drawbacks are small sample problem. Also the use of PE is sufficient to obtain satisfactory results. In [4] Asymmetric Bagging and Random Subspace based Support Vector Machine (ABRS-SVM) is present to solve the problems of SVM in image retrieval and over fitting problem. In [5], Navigation Pattern based Relevance Feedback (NPRF) achieve high efficiency and effectiveness with the large scale image data. Also reduces number of iterative feedbacks to produce refined search results. The iterative feedbacks are reduced substantially by using the navigation patterns discovered from the user query log. This NPRF approach is divided into two operations that is the online image retrieval and offline knowledge discovery. NPRF Search makes use of the discovered navigation patterns and three kinds of query refinement strategies such as Query Point Movement (QPM), Query Reweighting (QR), and Query Expansion (QEX). The query image is submitted to this system, and then the system first finds the most relevant images and returns it. This process is called initial feedback. Next, the positive samples picked up by the user is given to the image search phase including new feature weights, new query points and user's intention. Navigation patterns with three search strategies are included to find the desired images. For each user's browsing behaviors', offline operation for knowledge discovery is triggered to perform navigation pattern mining. The main drawbacks of this system are image retrieval in global feature space and results depends only on the navigation pattern of users. In [6], a new dimensionality reduction algorithm for relevance feedback in the content based image retrieval is called Biased Discriminative Euclidean Embedding (BDEE). The samples in the original dimensional ambient space is transformed to low level visual features to discover intrinsic coordinates of an image. BDEE models both the interclass geometry and interclass discrimination of each image. It does not ignore the manifold structure of samples. BDEE is a subspace learning method in which mapping vector is used to map high dimensional space to low dimensional space. In [7], Feature Line Embedding Biased Discriminant Analysis (FLE-BDA) is proposed for performance enhancement in relevance feedback scheme. It maximizing margin between relevant and irrelevant samples at local neighborhood so that relevant images and query image can be quite close, while irrelevant samples are far away from relevant samples. In this subspace learning method, find a linear transformation matrix from relevant or irrelevant images that is used in dimensionality reduction. The retrieval process includes 1) A query image is given as an input to the IR system. After calculating the similarity values, gallery images are ranked. 2) Users label the relevant or irrelevant images according to their preference. 3) Then user's' feedback is adopted to find a new transformation. 4) The gallery images are re-ranked to obtain the retrieval results in the next round. Two labels are assigned to the top ranking images according to users preference. Feedback with relevant or irrelevant labels represents users preference. The within-class scatter is calculated from the image samples with positive labels, while the between-class scatter is calculated from those with negative labels. Based on these assigned labels, the within-class and between-class weighted graphs are constructed for maximizing the margin of relevant and irrelevant samples. Then new distance between query and images are calculated. The advantages are dimensionality reduction, solve singular problem in the high dimensional space, increases generalization and robustness using Laplacian regularization. The disadvantage is computational complexity is very high due to the large scale dataset. In [8], Conjunctive Patches Subspace Learning (CPSL) method for learning an effective semantic subspace by exploiting the user historical feedback log data with the current data. CPSL effectively integrate the discriminative information of labeled log images, geometry information of labeled log images and weakly similar information of unlabeled images. For creating a reliable subspace, need to build different kinds of local patches for each image. Apart from other Relevance Feedback techniques, Collaborative Image Retrieval system integrates regular online RF schemes with an offline feedback log data. The CIR systems first collect RF information from user which can be stored in an RF

log database. If user feedback log data is unavailable then the CIR system performs exactly like RF based CBIR system. If the user RF information is available, the algorithm can effectively exploit the user feedback log data. The image retrieval can be done in less iteration than regular RF schemes with the help of the user historical feedback log data.

3. Proposed System:

The general algorithm of our approach is presented in Table 1 and then the detail of each function. This algorithm is applied for all users 'U'. The function Add(param1, param2), allows us to add the param2 into the param1. So, there no overwriting of the param1.

The algorithm begins with generating the relevant resources R' to a given tag, where $R' = \{r'_1, \ldots, r'_v\}$ the set of relevant resources and 'v' the number of relevant resources and $R' \subseteq R$,by using the function Add() in order to add each relevant resource into R'. The step interrogates the IndexFile (the output of the indexation step). When a request/query is made it is treated by the same analyser used to build the index and then used to find the corresponding term(s) in the index. It provides a list of resources matching the query. In our context, a query is considered as a tag throughout the rest of this paper, presenting the algorithm of generation resources relevant to a given tag $t_h \in T$ (see Table 2).

After generating relevant resources (R') according to a specific tag (t_h), a score is assigned to each resource according to the assigned tag. The purpose of using such score is to separate the most relevant resources related to a specific tag. This score is the result of a function of similarity which takes into consideration the resource (textual) and the tag. Many similarity functions exist in the literature such as the similarity function supported by Lucene. A predefined function of similarity which is a variant of the TF-IDF scoring model is choosen. The choice of such a model is due to the fact that TF-IDF is an efficient and simple algorithm for matching words in a tag to resources that are relevant to that tag. However, the main limitation of such a model is that it does not take into consideration the relations between words (e.g. synonyms). The similarity function is described through the formula (1) as follows:

$$score(q, r) = coord \cdot queryNorm(q) \cdot \sum \{tf(t \in r) \cdot \mathrm{idf}(t)^2 \cdot t.\ getBoost()(t, r)\}$$

The term 't' is the result of the resource indexation process. Each term 't' is associated with a resource 'r'.

Table 1:

The general algorithm of the interest detection approach for a specific user 'u'.

Input: N u , T nuj , IndexFile
// T nuj is the set of tags of the neighbours //

Output: I u
I u = ∅ , R′ =∅, R″ = ∅
1.For each nuj ∈ N u
2. R′ = GenerationResourcesRelevantToTag (T nuj , IndexFile)
3. R″ = Scoring(R′, T nuj)
4. Add (I u , SelectionRelevantTag(T, R″))
5 .End For
Return I_u

Table 2:

The algorithm of the generation of the resources relevant to each tag.

GenerationResourcesRelevantToTag (T_{nuj}, IndexFile)

Input: N_u, T_{nuj}, IndexFile

Output: R'// Set of the resources relevant to a each $t_h \in T_{nuj}$

$R' = \varnothing$

1. For each $t_h \in T$ do
2. Add (R', LuceneGeneration (t_h, IndexFile))
/* Generate List of resources R' relevant to the tag.*/

3. End For
4. Return R′

Predefined scoring function are described as follows:

- score (q , r) is the score affected to a specific resource r according to a specific query q.
- coord (q , r) is a score factor based on how many of the query q terms are found in the specified resource r.
- queryNorm(q) is a normalizing factor used to make scores between queries comparable.
- tf (t ∈ r): Term Frequency of the term t in the resource r. It is defined as the number of times term t appears in the currently scored resource r.
- idf(t): Inverse Document Frequency measure the importance of a term t in all the collection of resources.
- t . getBoost () is a search time boost of term t in the query q. The boost is 1.0 by default.
- *norm* (*t* , *r*) is a value of different boost and length factors: (i) Document boost sets a boost factor for hits on any field of the current resource. This value will be multiplied into the score of all hits on this resource. (ii) Field boost sets the boost factor hits on the current field. This value will be multiplied into the score of all hits on this field of a resource. (iii) lengthNorm(field): computed when the resource is added to the index in accordance with the number of tokens of this field in the resource, so that shorter fields contribute more to the score. The returning value is a normalization factor for hits on this field of this resource.

The scoring function will run according to the field content. This function provides a result of the top-k resources R'' relevant to the query q considered as a tag, where $R'' = \{r''_1, \ldots, r''_w\}$, the set of top-k relevant resources, where 'w' is the number of relevant resources and $R'' \subseteq R'$. The function Add() in order to add each relevant resource according to a tag into R'' is used. The scoring algorithm of the resources is described in Table 3.

Table 3:

The algorithm of scoring the resources for a given tag.

Scoring (R', T_{nuj})

Input: R', T_{nuj}

Output: R''// Set of the top $-k$ $r_{v'} \in R'$ relevant to $t_h \in T_{nuj}$

$R'' = \varnothing$

1. For each $r_{v'} \in R'$ do
2. For each $t_h \in T_{nuj}$
3. score[]=score ($r_{v'}$, t_h) // Lucene scoring function
4. End For
5. Add (R'', Top-k Generation($r_{v'}$, score[]))
6. End For
7. Return R″

Table 4:

The algorithm of selection of relevant tags.

SelectionRelevantTag(T_{nuj}, R'')

Input: T_{nuj}, R''

Output: I_u

$I_u=\varnothing$

1. For each $t_h \in T_{nuj}$ do
2. If($\exists r_{v''} \in R''$, $t_h \exists \langle U, T, R'' \rangle$)
3. Add (I_u, t_h). // Add the tag t into the set of the relevant interests of the user u.
4. End If
5. End For
6. Return Iu

The algorithm generates the set of relevant resources (R'') from the previous set (R') according to the specific tag (t_h) and the top-k resources having the higher score. For example, using a tag="math", one resource belonging to R'' is associated to the one with the title="IXLMath" and its URL="http://www.ixl.com/". After scoring the resources,we test if the resource tagged by 'q' exists in the top-k result provided by the scoring function. If it is the case, the tag 'q' is stated as relevant to the resource.

4. Results:

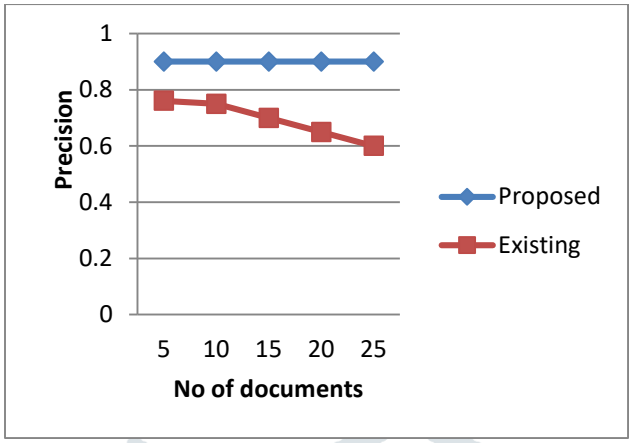

Figure 1.1: Precision values

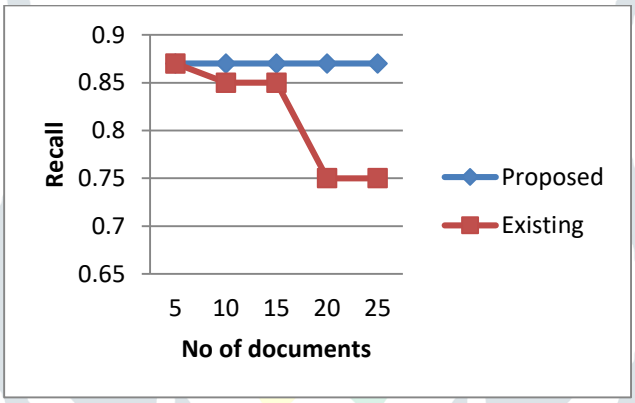

Figure 1.2: Recall value comparison

5. Conclusion :

The approach for detecting accurate user interests based on the social environment is proposed. The goal was to infer users interests from content of the tagged resources in order to figure out the tags really reflecting the thematic of the resources. The originality of the approach is based on the proposal of a new technique of interests detection by analysing the accuracy of the tagging behaviour of a user in order to figure out the tags which really reflect the content of the resources. So, these tags are somehow comprehensible and can avoid tags "ambiguity" usually associated to these social annotations. This is done through an indexation technique followed by an algorithm that score tags assigned to resources. The score reflects the relevance of the tag according to a resource. From this score, we have selected the most relevant resources (top-k). If the tag assigned by the user to a resource that is in the top-k, then the tag is considered an accurate interest. The experiment shows that the method provides a comprehensible set of interests. Consequently, this approach could be used for a purpose of adaptation (e.g. enrichment of the user profile, recommendation, etc.), since it provides a solution for detecting relevant user interests. The results have proved that the consideration of the tagged resources to detect the relevant user interests (our approach) is better than considering directly the tags assigned by the users (classical tag-based approach). In fact, the approach has treated the tag ambiguity and then, has provided better results. Future work can be focused on applying large datasets.

References:

[1] Ntalianis, K., Doulamis, A. D., Tsapatsoulis, N., & Mastorakis, N. E. (2018). "Social Relevance Feedback Based on Multimedia Content Power".

[2] Anelia Grigorova, Francesco G. B. De Natale, Charlie Dagli, and Thomas S. Huang, "Content-Based Image Retrieval by Feature Adaptation and Relevance Feedback" IEEE conference on Multimedia, Vol. 9, No. 6, pp. 1183-1192 ,Oct.2007.
[3] Mohammed Lamine Kherfi and Djemel Ziou, "Relevance Feedback for CBIR: A New Approach Based on Probabilistic Feature Weighting With Positive and Negative Examples, 2006.
[4] Dacheng Tao, Xiaoou Tang, Xuelong Li and Xindong Wu, "Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval," ,2006.
[5] I. Ja-Hwung Su, Wei Jyun Huang, Philip S. Yu, Fellow, and Vincent S. Tseng, "Navigation-Pattern Based Relevance Feedback For High Efficient Content-Based Image Retrieval," 2012.
[6] I. Wei Bian and Dacheng Tao, "Biased Discriminant Euclidean Embedding for Content-Based Image Retrieval,"2010.
[7] I. Yu-Chen Wang, Chin Chuan Han, Chen-Ta Hsieh, YingNong Chen, and Kuo-Chin Fan, "Biased Discriminant Analysis With Feature Line Embedding for Relevance Feedback-Based Image Retrieval," 2015.
[8] Lining Zhang, Lipo Wang, and Weisi Lin, "Baised subspace learning for SVM Relevance Feedback in Content-Based Image Retrieval,",2011.

[9] M. Amadasun and R. King, "Texural features corresponding to texural properties," IEEE Transaction on system, Man and Cybernatics,1989.

[10] S. C. Hoi, M. R. Lyu, and R. Jin, "A unified log-based relevance feedback scheme for image retrieval," IEEE Transaction on knowledge and data engineering, Vol. 18, No. 4,, April 2006.
[11] Tesic.J. and M. B, "Nearest neighbor search for relevance feedback. in computer vision and pattern recognition," IEEE Computer Society Conference, Volume 2,pp. 18–20, June 2003.