Personalized Facet Recommendation based on Conditional Random Fields

Yongquan Dong[a], Qiang Chu[b], Ping Ling[c]

School of Computer Science and Technology,

Jiangsu Normal University, Xuzhou 221116, China

[a]email: tomdyq@163.com, [b]email:785469372@qq.com, [c]lingicehan@163.com

Keywords: Facet Recommendation; Personalized Recommendation; Conditional Random Fields

Abstract. Faceted search is a kind of exploratory search, which is the complementary of keyword search. The current faceted search techniques display all facets. When there are plenty of facets, they are not available. Existing facet recommendation approaches mainly select facets according to the experts' experience or the statistics. They do not consider the user's interest which affects the recommendation result. Thus, this paper proposes a personalized facet recommendation approach based on conditional random fields. First, we use the user's query logs to build user profile and regard the facets in the logs which the user selects as his/her interested ones. Second, we select multiple kinds of features and use conditional random fields to construct the facet classification model. At last, when a user submits a query, the system selects all candidate facets and uses the facet classification model to predict their interest degrees. Then it sorts these facets from the highest degree to the lowest degree and shows top-k facets to the user. Experimental results validate the effectiveness of our approach.

Introduction

An increasing amount of information consists of a combination of both structured and unstructured data. For example, patent documents contain structured properties such as inventors, assignees, class codes, and filing date, as well as a body of unstructured text. Helpdesk tickets store not only structured data such as the tickets' originator, responsible party, and status, but also text describing the problem and its origin. Increasingly, enterprises want to run analytics[1] on text to extract additional valuable structured information such as chemical compounds used in a patent and products mentioned in a helpdesk ticket. As a result, the total number of structured properties in those data sets can be fairly large. Performing analysis on such data sets becomes challenging since a user may not know which properties to focus on. Ideally, a user would like to just type in some keywords into a system which would then guide him to areas of interest.

A promising approach for such mixed data is faceted search [2], which is widely used by e-commerce sites such as amazon.com. Instead of waiting for the user to create structured queries from scratch, a facet search system allows the use to progressively narrow down the choices by choosing from a list of facets. One key problem to building a faceted search system is to select which facets to make available to the user at any one time. This is especially important when the data is very large. Some systems show users all available facets. This approach can quickly overwhelm the users and lead to diminished user performance. Other systems such as eBay[3] present a manually chosen subset of facets to the user, and the facet-values are ranked based on their frequency. Other systems, such as Flamenco[4], simply present the first few facet-values in an alphabetized list. For systems with a large number of facets, manually selecting and maintaining a system of "blessed" facets may be too time consuming. Also, a predefined search system may not serve all users of the system adequately. Existing approaches do not consider the user's interests and cannot provide satisfied facets to the users.

In this paper, we propose a personalized facet recommendation approach based on conditional random fields. We cast the facet recommendation as a classification problem. First, we use the user's query logs to build user profile and regard the facets in the logs which the user selects as

his/her interested ones. Second, we select multiple kinds of features and use conditional random fields to construct the facet classification model. At last, when a user submits a query, the system selects all candidate facets and uses the facet classification model to predict their interest degrees. Then it sorts these facets from the highest degree to the lowest degree and shows top-k facets to the user. Experimental results show the effectiveness of our approach.

Our Approach

The Process of Our Approach. The process is shown as follows:

(1) Generating the user profile. We use the user's query log to generate his/her user profile. This does not need explicit user labeling. We also use the user's query log and regard the facets which the user selected as his/her interested ones.

(2) Building the facet classification model. We select plenty of facet content features and multiple kinds of similarity functions to compute the similarity between the user profile and facets content. And then we use the conditional random fields to build the facet classification model.

(3) Ranking the facets for a query. When a user submits a query, the system selects all candidate facets and uses the facet classification model to produce their interest degrees. Then it sorts these facts from the highest to lowest and returns top-k facets to the user.

Generating the user profile. User profile is used to represent user interest. In this paper, we use the information in the user's query logs to generate the user profile. The content of a facet is often represented as a document. The representation of a document is commonly based on vector space model. In this model, each document is represented by a feature vector in n-dimensional space and each dimension represents a distinct feature term extracted from the document. This representation is also the same for user profile that is defined based on the feature vector. User profile is often composed of two parts: a positive interest I_p^U which contains the facets selected by the user and a negative interest I_n^U which contains the facets not selected by the user. Each interest consists of a list of pairs of terms and their weights. The formula is as follows:

$$I_i^U = < w_j^i > \qquad (1)$$

where $i \in \{p,n\}$, w_j^i denotes the weight term t_j in I_i^U.

The TF-IDF weighting scheme [5] is used to assign the terms' weights. Based on TF-IDF, the term importance is proportional to the frequency of occurrence of each term in each document and inversely proportional to the total number of documents in a document collection in which the term occurs. It assumes that keywords appearing in fewer documents discriminate better than the ones appearing in more documents. So the weight of term t_j in I_i^U is then defined as follows:

$$w_j^i = tf_j^i \cdot \log \frac{N}{df_j^i} \qquad (2)$$

where N is the number of documents in the document collection, df_j^i is the document frequency of term t_j, and tf_j^i is the frequency of term t_j in I_i^U. Then m highest weighted terms are kept and normalized according to the information gain.

The user profile is denoted by P^U which is described as follows:

$$P^U = < I_p^U, I_n^U > \qquad (3)$$

Then we use the similar approach to model every facet content. And the content of the facets selected by the user is the positive interest facet IF_p^U and the others as the negative interest facet IF_n^U.

Building the facet classification model. We regard IF_p^U and IF_n^U as the annotated training set. Then we will use plenty of facet content features and multiple kinds of similarity function to measure the similarity between the training set and user profile and use these features to train a facet classification model which can predict a facet whether a positive facet or a negative facet with a probability. So in this section, we first introduce the linear-chain conditional random fields, then we give a description of the used features.

Linear-Chain CRF. A conditional random field is an undirected graphical model that defines a single exponential distribution over label sequences given a particular observation sequence. Let X

be a random variable over the observations to be labeled, and Y be a random variable over corresponding labels. All components Y_i of Y are assumed to range over a finite label alphabet y. In a discriminative framework, CRF construct a conditional model $p(Y \mid X)$ with a given set of features from paired observations and labels. Formally, the definition of CRF [6] is given subsequently:

DEFINITION 1. Let $G = (V, E)$ be an undirected graph such that $Y = \{y_v \mid v \in V\}$. Then (X, Y) is said to be a conditional random field if, when conditioned on *X*, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v \mid X, Y_{V-\{v\}}) = p(Y_v \mid X, Y_{N_v})$, where $V - \{v\}$ is the set of nodes in the graph except the node *v* and N_v is the set of neighbors of the node *v* in Graph *G*.

Thus, a CRF is a random field globally conditioned on the observations X. In theory the structure of graph *G* can be arbitrary. Many previous applications use the Linear-chain CRF, in which a first-order Markov assumption is made on the hidden variables. By the Hammersley-Clifford Theorem, the conditional distribution of the labels *y* given the observations *x* has the form,

$$p_\Lambda(y \mid x) = \frac{1}{Z(x)} \prod_{c \in C} \exp\left(\sum_k \lambda_k f_k(y_c, x_c) \right) \quad (4)$$

where $f_k(y_c, x_c)$ is a feature function of vertex and edge; λ_k is a learned weight associated with f_k, Λ is the weight set, $\Lambda = \{\lambda_k\}$, and $Z(x)$ is the normalization factor, also known as partition function, which has the form, $Z(x) = \sum_y \prod_{c \in C} \exp\left(\lambda_k f_k(y_c, x_c)\right)$

Training of CRF requires estimating the values of the weight set, Λ, which is usually done by maximizing the log-likelihood of a given training set. Popular training methods include generalized iterative scaling, conjugate-gradient and limited-memory quasi-Newton.

Inference in CRF is done by finding the most probable label sequence, $\hat{y}$, for an input sequence *x*, given the model in equation (4):

$$\hat{y} = \arg\max_y p_\Lambda(y \mid x) = \arg\max_y \sum_{c \in C} \sum_k \lambda_k f_k(y_c, x_c) \quad (5)$$

$\hat{y}$ can be searched by Viterbi algorithm using in HMM.

Feature Selection. We select multiple kinds of features which are shown as follows.

(1) Facet Content Features

a) Orthographic Features: average length of all facet content, all digits, including /, etc.

b)Keyword Features: Many words occur more frequently in some facets. These words (defined as keywords) can help to identify the categories of the facets. We automatically extract keywords which occur not fewer than 20 times from the training data.

c)Entity Features: time, date, location, person, etc.

(2) Similarity Features

We select multiple kinds of similarity functions to measure the similarity between the user profile and facet content. The similarity functions contain Jaccard[7], cosine[7], dice[7], etc.

Experiments

Dataset. In this paper, we use the DBLP dataset[8]. This dataset contains about 13,000 papers published in 26 venues(e.g. SIGMOD, VLDB, TODS, etc) in the past 30 years. It has 14 facets, including author, venue, time, location, number of authors per paper, and number of citations per paper. We ask five persons to browse this datasets and records the selected facets in query logs. Every person selects 100 facets.

Evaluation Criteria. To evaluate the performance, we define the Top-k accuracy, that is to say, it is computed by the ratio of the number of facets selected by user to the number of recommended Top-k facets.

Experiment Results. Table 1 shows the accuracy of top-1 to top-5 from 5 users and its average accuracy.

Table 1. Recommendation Performance

Top-k	User1(%)	User2(%)	User3(%)	User4(%)	User5(%)	Average(%)
Top-1	80.5	78.9	81.7	80.6	82.2	80.78
Top-2	90.2	89.4	91.0	90.8	92.4	90.76
Top-3	92.4	91.3	93.2	93.0	94.5	92.88
Top-4	98.6	97.6	1.0	98.8	99.5	98.9
Top-5	100.0	100.0	100.0	100.0	100.0	100.0

From Table 1, we can see that from Top-1 to Top-5, a gradual improvement in the accuracy of every user is obtained. The average Top-1 accuracy is 80.78%. And all the facets selected by the user are ranked among the top-5 recommendation result, which shows that our method has a good performance.

Conclusion

In this paper, we propose a personalized facet recommendation approach based on conditional random fields. First, we use the user's query logs to build user profile and regard the facets in the logs which the user selects as his/her interested ones. Second, we select multiple kinds of similarity features and use conditional random fields to construct the facet classification model. At last, when a user submits a query, the system selects all candidate facets and uses the facet classification model to predict their interest degrees. Then it sorts these facets from the highest degree to the lowest degree and shows top-k facets to the user. Experimental results validate the effectiveness of our approach.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No. 61100167 and 61105129, the Natural Science Foundation of Jiangsu Province, China under Grant No. BK2011204, Qing Lan Project, the National Training Program of Innovation and Entrepreneurship for Undergraduates under Grant No. 201310320052, the Practice Innovation Training Program Project for the Jiangsu College Students under Grant No. 201310320052Z and the Graduate Scientific Research and Innovation Project of Jiangsu Normal University under Grant No. 2013YYB130.

References

[1] Wisam Dakka, Dayal Rishabh, Ipeirotis P. Automatic Discovery of Useful Facet Terms[C]. SIGIR Faceted Search Workshop, 2006.
[2] Ping Wu, Yannis Sismanis, Berthold Reinwald. Towards Keyword-driven Analytical Processing[C]. Proceedings of the ACM SIGMOD International Conference on Management of Data, 2007.
[3] eBay. http://www.ebay.com/, 2014.
[4] Marti Hearst, Ame Elliott, Jennifer English, Rashmi Sinha, Kirsten Swearingen, Ka-Ping Yee. Finding The Flow in Web Site Search[J].Communications of the ACM, 2002,45(9):42-49.
[5] G. G. Chowdhury. Introduction to Modern Information Retrieval (3rd edition), Neal-Schuman Publishers, 2010.
[6] Lafferty John, McCallum Andrew, Pereira Fernando. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]. Proceedings of the International Conference on Machine Learning, 2001.
[7] William W. Cohen, Pradeep Ravikumar, Stephen E. Fienberg. A Comparison of String Distance Metrics for Name-matching Tasks[C]. Proceedings of the 2nd International Workshop on Information Integration on the Web, 2003.
[8]DBLP dataset. http://hpi.de/naumann/projects/repeatability/datasets/dblp-dataset.html, 2014.