An Interests Discovery Approach in Social Networks based on a Semantically Enriched Bayesian Network Model

Akram Al-Kouz and Sahin Albayrak
DAI-Labor, Technical University of Berlin, Ernst-Reuter-Platz 7, Berlin, Germany

Keywords: Interests Discovery, Bayesian Networks, Social Networks.

Abstract: Knowing the interests of users in Social Networking Systems becomes essential for User Modeling. Interests discovery from user's posts based on standard text classification techniques such as the Bag Of Words fails to catch the implicit relations between terms. We propose an approach that automatically generates an ordered list of candidate topics of interests given the text of the users' posts. The approach generate terms and segments, enriches them semantically from world knowledge, and creates a Bayesian Network to model the syntactic and semantic relations. After that it uses probabilistic inference to elect the list of candidate topics of interests which have the highest posterior probability given the explicit and implicit features in user's posts as observed evidences. A primitive evaluation has been conducted using manually annotated data set consisting of 40 Twitter users. The results showed that our approach outperforms the Bag Of Words technique, and that it has promising indications for effectively detecting interests of users in Social Networking Systems.

1 INTRODUCTION

The lexical meaning of interest is something that concerns, involves, draws the attention of, or arouses the curiosity of a person[1]. Thereby, discovering the interests of users in Social Networking Systems (SNS) could form a milestone in User Modeling (UM) (Al-Kouz and Albayrak, 2012). Personalized web applications and services such as recommender systems and personalized search engines depend on UM (Xuehua et al., 2005) (Gotardo et al., 2008). With the advent of SNS, personalized systems became more important. In such systems users become the producers of valuable information which can be used to enrich the UM (Kim et al., 2009). Users produce information in different forms, one of them is text messages referred it as Posts hereafter. Building UM in SNS is predominantly based on explicit data in profiles (Cataldi et al., 2010). Knowing the interests of a user from the contents of his Posts could enhance the UM.

When user publishes Posts in SNS he assumes the audience users in the other end is familiar with the subject, as a result user uses a few number of terms and phrases that have implicit semantic relations. Knowing the implicit encyclopedic semantics relations between terms and phrases can provide some clues about the topic of interest. The implicit semantics of a word or phrase is the vector of its encyclopedic associations in world knowledge such as Wikipedia[2] concepts (Al-Kouz et al., 2011). In knowledge world, concepts are semantically tied together by links which forms a graph of semantically related entities or concepts (Schonhofen, 2008).

Posts have special characteristics that have made Natural Language Processing (NLP) techniques not applicable to catch the implicit syntactic relations between terms (Moschitti and Basili, 2004). Implicit syntactic relations are the grammatical links between content words of a sentence which denote grammatical relations between nouns, verbs, adverbs and adjectives (Stevenson, 1998).

Usually user submits Posts at different points in time. Knowing the temporal factor can dramatically affect in catching the semantic relations between the contents of Posts. Posts with large time windows are suppose to be weakly related. On the other hand, Posts with small time windows could have strong semantic relations (Abe and Tsumoto, 2010).

Text classification is one of the common techniques to discover interest of users from their Posts (Gauch et al., 2007). Traditional text classification techniques based on Bag Of Words (BOW) perform well on large and rich with content documents, because the word occurrence is high, frequency of words is enough to

[1] http://dictionary.reference.com/browse/interest

[2] http://www.wikipedia.org

Al-Kouz A. and Albayrak S..
An Interests Discovery Approach in Social Networks based on a Semantically Enriched Bayesian Network Model.
DOI: 10.5220/0004172103000305
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 300-305
ISBN: 978-989-8565-29-7

capture the explicit semantics of terms. The explicit semantic of a word is the vector of its occurrence associations within the text (Tang et al., 2011). When dealing with Posts, the BOW based techniques will not perform well as they would have performed on larger text documents.

The Naive Bayes classifier is one of the simplest text classifiers based on BOW, in that it assumes all attributes of the text document are independent of each other given the context of the class, which is not correct in some real-world tasks. In fact, the Naive Bayes classifier could be considered as a Bayesian Network, in which the network structure is fixed and nodes can have only one parent class node (Acid et al., 2005). This model fails to deal with the previously mentioned implicit syntactic, explicit and implicit semantic, and temporal relations problems.

Bayesian Network is a graphical model for reasoning under uncertainty. It represents direct connections between nodes. These direct connections are often causal connections (Korb and Nicholson, 2010). Based on the causal implicit relation between the components of Posts, we believe that a proper Bayesian Network model can catch the explicit and implicit relations in users' Posts.

In this paper we propose an approach to discover the interests of a user from the contents of his Posts based on a semantically enriched Bayesian Network model. The proposed approach consisting of four phases. First is the features extraction and generation, in this phase we exploit the structure of the original Posts and external concepts from the WordNet[3] dictionary to get seed features. Generated seed features can express the implicit syntactic and explicit semantic relations between terms. The Second phase is the semantical enrichment of the seed features from world knowledge to crate new semantic entities that catch the implicit semantic relations between seed features. The third phase is building the Bayesian Network to represent the conditional dependencies between the seed features and the semantic entities. Seed features are represented as root nodes, semantic entities are represented as internal nodes, while the leaf nodes represent the categories extracted from world knowledge. Finally we use a probabilistic inference algorithm to compute the posterior probabilities of the leaf nodes to discover and rank interests topics of the user.

2 RELATED WORK

Al-kouz and Albayrak have proposed a semantic graph based approach to detect a user's interests from his Posts (Al-Kouz and Albayrak, 2012). In their approach they present Posts of a user as a semantic graph of related entities. Posts and comments from other users are represented as semantic social graph. In the semantic graph and the semantic social graph, nodes are semantically enriched entities from Freebase[4], edges are the semantic relations between those entities. The authors propose a Root-Path-Degree algorithm to prune the semantic graph and semantic social graph and to find the most popular subgraph that may infer the interests of the user. The proposed algorithm outperformed the Naive Bayes classifier in discovering interest of users. Its performance was complex in-terms of CPU time and memory space. In addition to that, it does not take into consideration the temporal relations between Posts, which has a reasonable impact on the meaning.

In (Ahmad et al., 2011) Ahmad and others proposed a new ranking algorithm called "SNPageRank" to find the interests model in the Friendfeed[5] social network. The proposed algorithm utilized updated version of the PageRank algorithm. Instead of web pages, people in SNS and the connections between them are used as hyper links, and the connections between nodes are weighted. This algorithm works fine in determining users with high contribution in the social networks without determining in which field they are interested or experts.

DeCampose et al. (de Campos and Romero, 2009) have proposed a Bayesian Network models for hierarchical text classification from a thesaurus. The proposed model represents each term in rich text documents as a root node and finds a relation to some descriptors in a thesaurus. Experimental results showed that the proposed model outperformed the baseline methods including Naive Bayes. Nevertheless, this model suffers from the limitations of thesaurus such as size and domain oriented nature. This model takes into consideration the explicit semantic problem, but it does not provide solutions to the implicit syntactic and semantic problems.

In contrast to all these works, our approach takes into consideration the implicit syntactic, explicit semantic, and implicit semantic problems. In addition, the proposed approach provides a pruning mechanism to reduce the number of nodes in the generated Bayesian Network. The temporal relation is represented as a node with diverging arcs to the root nodes of the consequence Posts. Moreover, it efficiently represents the parent nodes of the root and internal nodes

[3]http://wordnet.princeton.edu

[4]http://www.freebase.com

[5]http://friendfeed.com

to keep the size of the Bayesian Network as small as possible.

3 APPROACH

Discovering the interests of users in SNS based on their Posts could enhance the reliability of UM. The task of our proposed approach is to 1) extract and generate text features from users' Posts, 2) construct a semantically enriched Bayesian Network model. The output model is used to automatically map a user U to a target ranked list of Wikipedia Pages $P = \{P_1, P_2, ..., P_n\}$. Where P is a set of leaf nodes in the Bayesian Network, and items $P_1, P_2, ..., P_n$ are the different nodes representing topics of interest sorted by their posterior probability.

Therefore, the scope of our approach is automatic classification of users' Posts based on a semantically enriched Bayesian Network. There are several characteristics in this task which make it valuable: 1) each term in the Posts is a feature that need to be represented as node in our Bayesian Network. This leads to a dimensionality problem. 2) Terms could occur repeatedly in Posts. This is an explicit semantic problem that should be considered in our model. 3) Term has relations with surrounding terms. This relation need to be represented and referred later as an implicit syntactic problem. 4) Posts have temporal patterns. Temporal patterns can propagate the implicit semantic relations between the components of different Posts. 5) Terms and phrases usually have encyclopedic semantic relations. This is an implicit semantic problem. To overcome these problems, our approach is divided into four pipeline phases as following.

3.1 Features Extraction and Generation

In this phase we exploit the structure of the original Posts and external concepts from WordNet dictionary to extract and generate seed features. Seed features are wighted by term or phrase frequency. Generated seed features expressed the implicit syntactic and explicit semantic relations between terms and phrases. We took into consideration the dimensionality problem.

Posts in general have the characteristics of sparsity, highly focused, not domain specific, noisy, short in length, informal, multilingual, and grammatical error prone (Al-Kouz and Albayrak, 2012). These characteristics of Posts made it hard to apply standard NLP techniques to catch the implicit syntactic relations between words(Moschitti and Basili, 2004). When using the BOW model to represent Posts, it neglects the contextual information in them (Gimpel et al., 2011), which leads to uncertainty in classification. Timing pattern between Posts can affect the semantic relations between terms and phrases in different Posts. Therefore, explicit features need to be extracted and new implicit features need to be generated in a way that, the implicit syntactic, explicit and implicit, and temporal relations between terms and phrases in different posts could be caught.

3.1.1 Explicit Features Extraction

First of all, Posts are segmented based on punctuation marks. Next, we tokenized the segments. Then, a pre-processing mechanism applied on tokens to remove stop-words and extract the term feature seeds. After that we applied a Bi-gram tokenizer to extract the phrase feature seeds. For simplicity we assumed the maximum size of phrases to be two terms. Term feature seeds and phrase feature seeds are the Unigram and Bi-gram explicit features respectively that handle the implicit syntactic relations in one segment. The frequency of the term feature seeds and phrase feature seeds, represents the explicit semantic of seeds. The output of this step is a set of features (FS) which contains term feature seeds, phrase feature seeds, and segments. Segment is considered as the child of its feature seeds. The explicit semantic of the feature seeds is used as the prior probability in a later phase, and its relation to child segment represents a causal relation between nodes in our Bayesian Network.

3.1.2 Implicit Features Generation

Seed features are validated and enriched using WordNet dictionary. Each seed feature is checked against WordNet. The seed features existed in WordNet are confirmed as valid seed features, while others are neglected. Validation process ensures the dimensionality reduction by pruning the seed features that could be noise. Valid seed features enriched with its synonyms from WordNet. I addition, the inverted phrase seed features are checked, and if some match in WordNet is found, then the inverse considered as generated seed feature as well. The WordNet enrichment process generates new features that can enhance the seed connectivity with other nodes in Bayesian Network construction phases. All the generated features are added to the FS which was generated from the previous step.

3.1.3 Implicit Temporal Relations

Posts have different time stamps. Timing pattern between Posts can affect the semantic relations between terms and phrases in different Posts. Time windows between Posts calculated and considered as temporal

features. Each two successive Posts generate a temporal feature. Temporal features are added to the FS with the time stamp of its successive Posts. These time stamps will be used to calculate the prior probabilities of temporal nodes in Bayesian Network construction phase.

3.2 Semantic Enrichment

The semantical enrichment of the seed features from Wikipedia creates new semantic entities that will be represented as internal nodes in the Bayesian Network. These nodes catch the implicit semantic relations between seed features. Knowing the implicit encyclopedic semantic relations between different features in the FS can add new clues about the topic of interest. Wikipedia has been recognized as a promising lexical semantic resource. We utilized a recent publicly available dump of Wikipedia to match seed features to Wikipedia pages and find the semantic relations between pages. The process of matching features to Wikipedia pages is handled by using the feature as a query term to search the local Wikipedia dump for a disambiguated page that matches the query criteria. Once the disambiguated page is retrieved, we use the page title as an identification label, the "Wiki Page Feature" as the node type, and the first section text as its description. The page is added to the FS for later usage in the Bayesian Network construction phase.

If the result of the query is an ambiguous set of potential Wikipedia pages, then a disambiguation problem encountered. We utilized some entity disambiguation methods to solve this problem. More specifically, given a set of candidate pages from Wikipedia, we execute a search on index fields storing page titles, redirect titles, and name variants. We implement a weighted search to give high weights to the exact title matches or matches with minimum edit distance. The ambiguous page with the highest weight is added to the FS.

To find the semantic relations between the "Wiki page Features" entries in FS, we utilize the InfoBox[6] and the first section links. For each "Wiki Page Feature" entry we parsed the InfoBox and the first section links only to keep the performance and reduce the noise. The links of target pages considered as semantic enrichment that can catch the semantic relations between"Wiki page Features" entries. Each semantic enrichment page is added to the FS, and weighted by the frequency of its appearance in the InfoBox plus one times the frequency of its appearance in the first section plus one. All entries generated from the semantical enrichment phase are considered as semantic features.

[6]http://en.wikipedia.org/wiki/Help:Infobox

3.3 Bayesian Network Construction

The third phase in our pipeline is ti build the Bayesian Network to represent the conditional dependencies between the temporal features, the seed features and the semantic features. Temporal features are represented as root nodes. Seed features and semantic features are represented as internal nodes, while the leaf nodes represent the categories extracted from Wikipedia.

Bayesian Network is a graphical model for reasoning under uncertainty defined as a pair $B = (G,P)$. Where $G = (V(G), A(G))$ is an acyclic directed graph with set of nodes $V(G) = X_1, X_2, ..., X_n$ represent variables, and a set of arcs $A(G) \subseteq V(G) * V(G)$ represent direct connections between nodes. These direct connections are often causal connections (Korb and Nicholson, 2010).

The construction of a Bayesian Network involves three major steps. First, we must decide on the set of relevant nodes and their possible values, as in subsection 3.1. Next, we must build the network structure by connecting nodes into an acyclic directed graph. Finally, we must define the Conditional Probabilistic Table (CPT) for each network node (Darwiche, 2009).

3.3.1 Bayesian Network Structure

Temporal features should generate the top level nodes in our Bayesian Network. We retrieve all temporal features from the FS. Each temporal feature is represented by a node with the time stamp as its id, and the node type is "Temporal". Each temporal node should has only two children to represent its two successive Posts. Each child is represented by a virtual node, "Left Virtual" and "Right Virtual" as shown in Figure 1. The virtual child node is the parent node of all seed feature nodes contained in its Post.

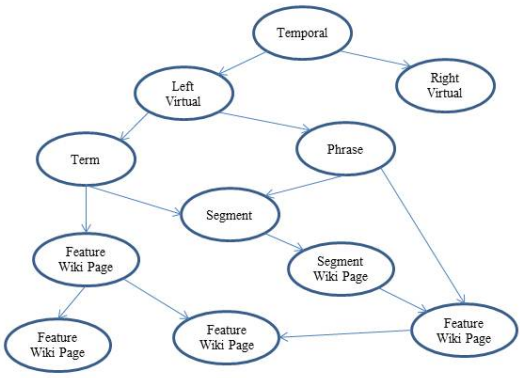

Figure 1: The Suggested Bayesian Network model.

We traverse the FS searching for segment entries, for each segment we add a new representing node to the

Bayesian Network. The added node has the segment text as node id and "Segment Text" as node type. After that, we try to retrieve the matching Wikipedia page. If it is exist, we add a new node to the Bayesian Network, this node has the page title as node id and "Segment Wiki Page" as node type. Further more, we create an arc connection outwards from "Segment Text" node to "Segment Wiki Page" node, because "Segment Text" node is the causal of "segment Wiki Page" node. Semantically connected Wikipedia pages to our "Segment Wiki Page" are retrieved and added as children nodes.

Parent nodes of each segment node need to be represented in our Bayesian Network. We retrieve the term seed features and the phrase seed features of the segment. Then we add a new node to represent the seed feature. Its id is the seed feature text and its type is "Term" or "Phrase" depending on the seed feature. The newly added nodes are connected to their "Segment Text" node as parent nodes. For each "Term" or "Phrase" seed feature node we retrieve its "Wiki Page Feature", then we add it to the Bayesian Network as child node. Node id is page title and node type is "Wiki Page Feature". Further more, every "Wiki Page Feature" connected to our current "Wiki Page Feature" is added to the Bayesian Network as child node, because it is connected semantically to current node by outlinks which implies casual relationship from current node to target nodes.

3.3.2 Conditional Probabilistic Tables

The CPTs of root nodes are easy to be calculated. Hence, they have only two states exist or not exist, and their probabilities are unconditional on any parent nodes. In our Bayesian Network model, root nodes are the Temporal nodes. The probability of Temporal node T is calculated in equation 1. In which $T1$ and $T2$ are the time stamps of the successive Posts. This equation expresses the temporal relations between Posts. It utilizes the assumption that Posts with small time window suppose to have strong semantic correlation. After that, we calculate the CPTs of each virtual node X in equation 2, Where $Pa(X)$ is the set of parent nodes of the current virtual node.

$$P(T) = \exp - |T1 - T2| \tag{1}$$

$$p(X) = \sum_{Pa(X)} p(X|Pa(X))P(Pa(x))) \tag{2}$$

The third level of CTPs is the CPTs of Term and Phrase nodes. The probability is calculated by dividing the node frequency calculated in 3.1.1 by the total number of Term and Phrase nodes in our Bayesian Network. Segment Text node is a child node of some Term and Phrase nodes, The same equation 3 is used to calculate the CPT of this node. Where $Pa(X)$ is the set of Term and Phrase parent nodes.

For the CPTs of the Wiki Page Feature nodes and Segment Wiki Page nodes, we treated them in the same manner. They will be referred as WP. Given its parent seed feature and Segment nodes its probability represented as $P(WP|Pa(WP)$ and calculated using the canonical method in equation 3. Where $parents$ is the set of Term parent, Phrase parent and WP nodes. Here, $w((P),WP)$ is the weight assigned to each parent p in $parents$ referring to this WP. The weight $w(p,WP)$ is calculated in equation 4.

$$P(WP|Pa(WP) = \sum_{p \in parents} w(p,WP) \tag{3}$$

$$w(p,WP) = \begin{cases} TF * IDF & if\ p = Term \\ & \vee Phrase \\ & \vee Segment \\ LW & if\ p = WP \end{cases} \tag{4}$$

Where TF is the function of node frequency of this p over the total Term, Phrase and Segment nodes frequency. IDF is the log of the number of WP in our Bayesian Network divided by the number of WP that contain this parent p. The LW is a function to calculate the probability of the parent node p to be in both the InfoBox and the first section links of WP.

One of the key issues arises here is the potentially large size of CPTs. To solve this problem, we pruned the Bayesian Network by removing all the nodes that have no connectivity with other nodes except their parents. This will reduce the total number of nodes in our Bayesian Network and ensures to have the minimum number of parents for each node. Reducing the number of parents is the main factor in reducing the size of CPTs. Hence, the CPT size of a specific node is X^{n+1}, where n is the number of parents, and X is the number of variable states.

3.4 Probabilistic Inference

In the last phase we use probabilistic inference algorithms to compute the posterior probabilities of the leaf nodes in order to discover and rank interest topics of user. The procedure used to map a given user U to a Wikipedia category C is as follows: first in the Bayesian Network, we instantiate the Term, Phrase, and Segment nodes corresponding to the words appearing in the Posts of U as observed and the remaining nodes as not observed. Let u be such a configuration of the Term, Phrase and Segment nodes. Next, we propagate this information through the network, and

compute the posterior probabilities of the Wiki Page feature nodes as $p(WP|Pa(WP))$. Finally, the ordered list of Wiki Page feature nodes with maximum posterior probabilities is used to map the user U to his topic of interest.

Further research needs to be conducted to investigate and apply an efficient inference algorithm, because different algorithms are suited to different network structures and performance requirements (Korb and Nicholson, 2010). Primitive test using different standard inference algorithms showed the reliability of our proposed approach.

4 CONCLUSIONS

Interest detection in Social Networks has attracted much attention recently. In this paper, we addressed the problem of mapping Users to topics of interest. Differently from previous work using the BOW based text classification techniques, we proposed a technique based on a Bayesian Network model to represent the implicit syntactic, explicit semantic, implicit semantic and temporal relations between the Posts of a user. According to the primitive experimental results, our proposed approach showed promising indications. In future we would like to investigate different inference algorithms to calculate the posterior probability of the candidate topics of interests.

REFERENCES

Abe, H. and Tsumoto, S. (2010). Text categorization with considering temporal patterns of term usages. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pages 800–807, Washington, DC, USA. IEEE Computer Society.

Acid, S., de Campos, L. M., and Castellano, J. G. (2005). Learning bayesian network classifiers: Searching in a space of partially directed acyclic graphs. *Machine Learning*, 59:213–235. 10.1007/s10994-005-0473-4.

Ahmad, K., Amin, O., and Farzad, F. (2011). Expert finding on social network with link analysis approach. In *19th Iranian Conference on Electrical Engineering (ICEE), 2011*, ICEE 2011, Iran.

Al-Kouz, A. and Albayrak, S. (2012). An interests discovery approach in social networks based on semantically enriched graphs. In *Proceedings of The 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM2012, Istanbul, Turkey.

Al-Kouz, A., Luca, E. W. D., and Albayrak, S. (2011). Latent semantic social graph model for expert discovery in facebook. In *Proceedings of 11th International Conference on Innovative Internet Community Systems*, I2CS '11, Berlin, Germany.

Cataldi, Mario, Caro, D., Luigi, Schifanella, and Claudio (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA. ACM.

Darwiche, P. A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, New York, NY, USA, 1st edition.

de Campos, L. M. and Romero, A. E. (2009). Bayesian network models for hierarchical text classification from a thesaurus. *Int. J. Approx. Reasoning*, 50(7):932–944.

Gauch, S., Speretta, M., Chandramouli, A., and Micarelli, A. (2007). User Profiles for Personalized Information Access The Adaptive Web. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, chapter 2, pages 54–89. Springer Berlin / Heidelberg, Berlin, Heidelberg.

Gimpel, K., Schneider, N., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments.

Gotardo, R., Teixeira, C., and Zorzo, S. (2008). Ip2 model - content recommendation in web-based educational systems using user's interests and preferences and resources' popularity. In *Proceedings of 32nd Annual IEEE International Computer Software and Applications*, COMPSAC '08, Turku, Finland.

Kim, J.-T., Lee, J.-H., Moon, J.-Y., Lee, H.-K., and Paik, E.-H. (2009). Provision of the social media service framework based on the locality/sociality relations. In *IEEE 13th International Symposium on Consumer Electronics*, ISCE '09, Dajeon, South Korea. IEEE.

Korb, K. B. and Nicholson, A. E. (2010). *Bayesian Artificial Intelligence, Second Edition*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition.

Moschitti, R. and Basili, R. (2004). Complex linguistic features for text classification: a comprehensive study. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR*, pages 181–196. Springer Verlag.

Schonhofen, P. (2008). Annotating documents by wikipedia concepts. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, volume 1, pages 461 –467.

Stevenson, M. (1998). Extracting syntactic relations using heuristics. In *ESSLLI98 - Workshop on Automated Acquisition of Syntax and Parsing*, pages 248–256.

Tang, J., Wang, T., Lu, Q., Wang, J., and Li, W. (2011). A wikipedia based semantic graph model for topic tracking in blogosphere. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI '11, Barcelona, Spain.

Xuehua, S., Bin, T., and ChengXiang, Z. (2005). Implicit user modeling for personalized search. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, New York, NY, USA. ACM.