Neurocomputing 172 (2016) 27–37

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Personalized search for social media via dominating verbal context[☆]

Haoran Xie [a], Xiaodong Li [b], Tao Wang [c], Li Chen [d], Ke Li [b], Fu Lee Wang [a], Yi Cai [c,*], Qing Li [b,e], Huaqing Min [c]

[a] *Caritas Institute of Higher Education, Hong Kong SAR, China*
[b] *Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China*
[c] *School of Software Engineering, South China University of Technology, Guangzhou, China*
[d] *Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China*
[e] *Multimedia Software Engineering Research Centre, City University of Hong Kong, Hong Kong SAR, China*

ARTICLE INFO

Article history:
Received 1 December 2013
Received in revised form
8 December 2014
Accepted 17 December 2014
Available online 19 May 2015

Keywords:
Dominating set
Context
Folksonomy
Personalized search
Collaborative tagging

ABSTRACT

With the rapid development of Web 2.0 communities, there has been a tremendous increase in user-generated content. Confronting such a vast volume of resources in collaborative tagging systems, users require a novel method for fast exploring and indexing so as to find their desired data. To this end, contextual information is indispensable and critical in understanding user preferences and intentions. In sociolinguistics, context can be classified as *verbal context* and *social context*. Compared with *verbal* context, *social* context requires not only domain knowledge to build pre-defined contextual attributes but also additional user data. However, to the best of our knowledge, no research has addressed the issue of irrelevant contextual factors for the *verbal context* model. To bridge this gap, the dominating set obtained from *verbal context* is proposed in this paper. We present (i) the *verbal context graph* to model contents and interrelationships of *verbal context* in folksonomy and thus capture the user intention; (ii) a method of discovering dominating set that provides a good balance of essentiality and integrality to de-emphasize irrelevant contextual factors and to keep the major characteristics of the *verbal context graph*; and (iii) a revised ranking method for measuring the relevance of a resource to an issued query, a discovered context and an extracted user profile. The experimental results obtained for a public dataset illustrate that the proposed method is more effective than existing baseline approaches.

1. Introduction

With the rapid development of Web 2.0 communities, a tremendous increase in user-generated content has been witnessed in recent years. A typical example in Web 2.0 is collaborative tagging systems (known as folksonomy), which allows users not only to upload and share their generated contents such as photographs, videos or blogs, but also to express their subjective feelings on interesting content with semantic-rich tags. Confronting such a vast volume of resources in collaborative tagging systems, users require a novel method for fast exploring and indexing so as to find their desired data. To this end, contextual information is indispensable and critical in understanding user preferences and intentions.

In sociolinguistics, context can be classified as *social contexts* and *verbal contexts*. A *social context* refers to the social identity being construed and displayed in text and talk by language users [11]. Social contexts are often defined by objective social variables, such as those of class, gender or race. Therefore, in IR (information retrieval) research communities, *social contexts* are modeled as pre-defined attributes accordingly, and each contextual attribute has a certain value. As illustrated in Table 1, an example of a *social context* model contains three contextual attributes (mood, weather and time) and their respective values (e.g., mood has values such as 'sad' or 'happy'). The notion of verbal context, which is generally regarded as the text or talk surrounding an expression [13], is usually interpreted as the collection (or structure) of previous queries and/or actions (e.g., click-through data) in a search task by the user [9,30].

The problem of importing irrelevant contextual factors [34], which reduces the effectiveness and efficiency, is a great challenge for both *social* and *verbal* contextual approaches. To deal with this problem, the feature selection [14] has been adopted to filter trivial attributes in *social context* models [28,34]. Compared with *verbal* context, *social* context requires not only domain knowledge to build pre-defined contextual attributes but also additional user data. However, existing approaches for feature selection cannot be employed in *verbal context* models, because the contextual variables are independent of each other in a *social context* while contextual factors in a *verbal context* are highly intercorrelated. Moreover, to the best of our knowledge, no research has addressed

[☆] This paper is an extended version of our previous conference paper [16].
[*] Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.12.109
0925-2312/© 2015 Elsevier B.V. All rights reserved.

Table 1
Example of a social context model.

Contextual attribute	Contextual values
Mood	Sad, happy, scared, angry, neutral, etc.
Weather	Sunny, rainy, snowy, cloudy, stormy, etc.
Time	Morning, afternoon, evening, night, etc.

the issue of irrelevant contextual factors for the *verbal context* model. To bridge this gap, this paper proposes finding important factors (referred to as the dominating set) from the *verbal context*. We consider that dominating set should reflect characteristics of the original context more precisely and assume that a dominating set is a good candidate (possibly even better than the *verbal context*) with which to represent the *verbal context*. The following are the main contributions made in this paper:

- We present the verbal contextual graph to model *verbal context* in the collaborative tagging systems (folksonomy) and thus capture the user intention and facilitate a personalized search.
- In contrast with importance-based approaches (e.g., PageRank [5], HITS [17]), a method of discovering a dominating set, which has a good balance of essentiality and integrality, is proposed for the *verbal context graph* to prune the irrelevant contextual factors while keeping the major characteristics.
- We propose a revised ranking method for measuring the relevance of a resource to an issued query, a discovered context and an extracted user profile.
- We conduct experiments on a public dataset, and validate the effectiveness and efficiency of our proposed approach by comparing it with baselines.

The remaining sections of this paper are organized as follows. Context modeling approaches in IR and personalized searches in folksonomy are reviewed in Section 2. In Section 3, the contextual graph for *verbal context* in folksonomy and the dominating set are formulated. Experiments are conducted and reported in Section 4. Section 5 summarizes our work and discusses future research plans.

2. Related work

2.1. Context modeling

The main strategies used to incorporate the context the recommendation and IR systems can be categorized as pre-filtering, post-filtering and contextual modeling [2]. The three different strategies were experimentally investigated by Panniello et al. [23]. From the perspective of context, context modeling can be categorized into the aforementioned *verbal* and *social* contexts, which are the focus of this review and explained as follows.

Social context model: *Social* contextual approaches model the context as predefined attributes, with each contextual attribute having certain values. Adomavicius et al. proposed multiple dimensions that represent the context, with each dimension being a subset of a Cartesian product of the predefined attributes [1]. Moreover, *social context* has been widely adopted in domain-specific applications. Kosir et al. built a context movie database (LDOS-CoMoDa) and defined the context as a set of contextual variables (time, location, mood) related to the movies and audiences [18]. Wang et al. [29] presented context by five different daily activities to facilitate a context-aware mobile music recommendation. In a restaurant recommender system, Vargas et al. defined 46 contextual variables, and adopted a feature selection method to obtain relevant variables [28].

Verbal context model: *Verbal* contextual approaches focus on contextual elements such as past queries or click-through data of the user in the same task. White et al. represented context by the ODP (Open Directory Project) categories of web pages previously visited URLs by the user within a task session to facilitate the prediction of user interest [30]. Additionally, past click-through data were collected as context and incorporated into a conditional random field model to address the problem of query classification [9]. By mining latent concept patterns in a search log, Liao et al. presented a context-aware query suggestion model [19]. Moreover, Cantador et al. depicted context using ontological terms and their semantic relationships for historic data [8].

2.2. Personalized search in folksonomy

Folksonomy-based systems (e.g., Flickr[1] and Delicious[2]) have gained great popularity in the Web 2.0 era. Confronting the increasingly large volume of user-generated data, users require effective and efficient personalized approaches to find interesting resources in such applications according to their preferences. In folksonomy, the personalized search has been mainly supported and facilitated by the tag-based user profile [22]. More recently, there have been works on the paradigms of the construction of user and resource profiles. Xu et al. adopted Term Frequency/Inverse Document Frequency (TF-IDF) from IR to build tag-based profiles [33], while Best Matching 25 (BM 25) and the Normalized Tag Frequency (NTF) were proposed by Vallet et al. [27] and Cai et al. [7] respectively. In addition, nested context modeling was presented to generate a context-aware personalized search result in our previous work [32]. In any event, domain-specific applications, specifically multimedia objects [21], social bookmarks [10] and web pages [3], have greatly beneficial from the use of social tags.

3. Formulation

In this section, we formulate the research problem and discuss the proposed model. Generally, the problem of a context-aware personalized search in folksonomy can be formally defined as a mapping function θ:

$$\theta : R \times U \times Q \times C \rightarrow S \in [0,1], \quad (1)$$

where R, U, Q and C are the sets of resources, users, queries and contexts respectively, and S is a set of relevant ranking scores of the resource set in the range $[0,1]$. Different from the existing formulation of a personalized search problem in folksonomy [7,27,33], we incorporate the context set C into the mapping function θ so as to make the search result context-aware.

In the following subsections, we firstly introduce how to extract user and resource profiles (i.e., representations of the users and resources) in the folksonomy. We then formulate the context model in detail. Next, the algorithm used to find the dominating set is proposed. Finally, we revisit the mapping function θ, and show how the ranking score of the context-aware personalized search is obtained.

3.1. User and resource profiles extraction

Tag-based user and resource profiles have been adopted to facilitate personalization in the folksonomy. The core idea is that user-annotated tags, which reflect both subjective user preferences and objective resource contents, are a valuable and useful

[1] http://www.flickr.com
[2] http://www.delicious.com/

information source for depiction of users and resources. Therefore, it is natural to extract tag-based profiles in representing users and resources from the *folksonomy*, which is formally defined as follows.

Definition 1. A *folksonomy*, denoted by F, is a tuple of four elements:

$$F = (U, R, T, P),$$

where U, R and T are the sets of users, resources and tags respectively, and P is a set of relations between the three sets and satisfies $P \subseteq U \times R \times T$.

There are various paradigms (e.g., TF-IUF/IRF, BM 25 and NTF) that extract user and resource profiles from the above-defined *folksonomy F*. In this research, we employ NTF to construct tag-based user and resource profiles as it was proven to be more rationale and effective than other paradigms in our previous work [7]. Specifically speaking, a *user profile*, which is a vector of the tag-value pair, can be extracted from the folksonomy as defined below.

Definition 2. Let $\{t_1^i, \ldots, t_n^i\}$ and $\{\tau_1^i, \ldots, \tau_n^i\}$ be tags used and their degrees of relevance (or preference) by user u_i respectively. The *user profile* is represented by a vector $\overrightarrow{u_i}$:

$$\overrightarrow{u_i} = (t_1^i : \tau_1^i, t_2^i : \tau_2^i, \ldots, t_n^i : \tau_n^i).$$

The degree of relevance τ_n^i is obtained via the NTF paradigm as

$$\tau_n^i = \frac{K_n^i}{K^i}, \tag{2}$$

where K_n^i is the frequency that user u_i uses t_n^i to annotate resources, and K^i is the total number of resources tagged by that user. A larger value of τ_n^i indicates that the tag t_n^i is more relevant to (preferred by) user u_i.

Similarly, the *resource profile* can be defined as a tag-value pair vector and derived from the folksonomy via the NTF paradigm which is represented as follows.

Definition 3. Let $\{t_1^x, \ldots, t_m^x\}$ and $\{\upsilon_1^x, \ldots, \upsilon_m^x\}$ be tags and their degrees of relevance to (or representation of) resource r_x, respectively. The *resource profile* is denoted by the vector $\overrightarrow{r_x}$:

$$\overrightarrow{r_x} = (t_1^x : \upsilon_1^x, t_2^x : \upsilon_2^x, \ldots, t_m^x : \upsilon_m^x).$$

Accordingly, the degree of relevance υ_m^x is acquired via the NTF paradigm as

$$\upsilon_m^x = \frac{L_m^x}{L^x}, \tag{3}$$

where L_m^x is the number of users choosing tag t_m^x to annotate resource r_x, and L^x is the total number of users tagging that resource. A larger υ_m^x value indicates that tag t_m^x is more relevant to (representative of) resource r_x. Since other paradigms such as those of TF, TF-IUF/IRF and BM 25 for constructing user and resource profiles can be included in this framework, we will study their effects on the personalized search performance in the experiment described later.

Note that tag-based user and resource profiles (as defined in Definitions 2 and 3) from folksonmy (Definition 1) can be extracted off-line [33] to avoid costly re-computation when a new relation is added (i.e., when a user annotates a resource). In the next subsection, we further discuss how the query (Q) and context (C) are obtained (as expressed in Eq. (1)), since the user (U) and resource (R) have been extracted from the folksonomy.

3.2. Verbal contextual graph construction

The notion of (verbal) context is generally regarded as the surrounding text or talk surrounding an expression in linguistic study [13]. In the scenario of a personalized search, the expression normally refers to a *query* issued by the user. Formally, we denote a *query* as a term-value pair vector as follows.

Definition 4. Let $\{t_1^i, \ldots, t_l^i\}$ and $\{\varsigma_1^i, \ldots, \varsigma_l^i\}$ be query terms (or tags) and their degrees of importance to the query result specified by user u_i. The *query* is in the form of a vector $\overrightarrow{q_i}$:

$$\overrightarrow{q_i} = (t_1^i : \varsigma_1^i, t_2^i : \varsigma_2^i, \ldots, t_l^i : \varsigma_l^i).$$

The degree of importance ς_l^i takes a value of 1 by default, except when the issuer specifies the value to highlight (or attenuate) terms/tags in the query.

The query context, which is the text or talk surrounding an expression (query), can be interpreted as “the queries previously issued by the user in a search task” in a personalized search. For example, in a system that searches for cooking recipes, a user issues a query “beef” and the corresponding verbal context could be terms like “spicy, braised” from previous queries within the same conversation; another example from the Movielens[3] dataset is that the verbal context for “kung fu” is a set of tags such as “bloody action movies”.

Intuitively, it is straightforward to define context as “a bag of words” to collect all terms from previous queries in a search task. However, the main reason why the feature selection method cannot be applied to *verbal context* directly is that the elements of the surrounding text are highly correlated with each other. To tackle this problem, we present a *verbal contextual graph* to model not only the contextual elements but also their relationships of surrounding text (verbal context). To be specific, a *verbal contextual graph* is formally denoted below by a tuple of two elements.

Definition 5. A *verbal contextual graph* for a query q_i, denoted by G_i, is an undirected graph:

$$G_i = \{V_i, E_i\}$$

$$V_i = \left\{ t_a \mid t_a \in \bigcup_{k=1}^{i-1} q_k \right\}$$

$$E_i \subseteq V_i \times V_i,$$

where E_i is the edge set and V_i is the vertex set, which is consisted by terms/tags in the collections of queries issued before q_i by the user within a task session. The boundary of a task session can be demarcated using the 30-minute threshold strategy that is widely used in Web log analysis [31].

Note that the query terms (tags) are not included in this definition of a verbal contextual graph. We refer to this contextual graph as a *query-excluded contextual graph* to distinguish it from a *query-included contextual graph*, which integrates all query terms of current query q_i in the vertex set V_i (i.e., $V_i = \{t_a \mid t_a \in \bigcup_{k=1}^{i} q_k\}$). In the experiment described later, we compare the effects of these two types of contextual graph on all context-aware approaches.

As term (tag) vertices in a contextual graph may have different frequency, we assume that these tags, which have higher frequencies, have higher possibilities of reflecting the user's intentions. According to this assumption, we adopt the normalized tag

[3] http://www.grouplens.org/node/73

frequency as the weight of a vertex in the context graph:

$$u(t_a)=\frac{f(t_a)}{\sum_{\forall i} f(t_i)}, \tag{4}$$

where $f(t_a)$ is the frequency of t_a for all queries within the same session, and $\sum_{\forall i} f(t_i)$ is the total number of frequency for all terms. The intercorrelation between elements (terms) in a *verbal contextual graph* is modeled as the weights of edges. In this research, we adopt WordNet[4] and Lin's similarity [20] to capture the semantic relationship between the terms. The main reason for selecting WordNet is that the latent semantic relationships (e.g., the "is-a" relationship) are measured. Specifically, the weight of an edge in E_i is calculated as

$$w(e_{ab})=Sim(t_a,t_b)=\frac{2\times\log(p(lso(\tilde{t_a},\tilde{t_b}))}{\log(p(\tilde{t_a}))+\log(p(\tilde{t_b}))}, \tag{5}$$

where e_{ab} is the edge between vertices t_a and t_b, $e_{ab}\in E_i$ and $t_a, t_b\in V_i$, $\tilde{t_a}$ and $\tilde{t_b}$ are the synsets for terms t_a and t_b respectively, $lso(\tilde{t_a},\tilde{t_b})$ is the lowest super-ordinate to the information content of $\tilde{t_a}$ and $\tilde{t_b}$, and $p(\tilde{t_a})$ is the probability of matching an instance of the synset $\tilde{t_a}$ in the corpus. A larger value of $w(e_{ab})\in[0,1]$ indicates greater correlation of the vertex pairs in terms of their semantics.

To tackle the problem of importing irrelevant contextual factors in verbal context, we can prune the irrelevant parts (or highlight the core parts) of the *verbal contextual graph* constructed above. A straightforward approach is to discover the core vertices in a *verbal contextual graph* using the PageRank algorithm [5] or HITS algorithm [17]. However, we argue that the core vertices in the *verbal contextual graph* should have the following two properties.

- *Essentiality*: The core vertices should be more essential than the pruned vertices to reflect the nature (or main contents) of the *verbal contextual graph*.
- *Integrality*: The core vertices should contain the semantics of the pruned vertices and keep the complete semantics of the *verbal contextual graph*.

If we adopt PageRank or HITS, the property of essentiality will be ensured after the pruning step. However, these algorithms may not provide core vertices that have the property of integrality as they only consider the degree of importance of the vertices from link analysis. To retain a good balance of essentiality and integrality, we borrow the idea of the dominating set in graph theory to find core vertices. The method comprises two sub-processes: (i) we convert a weighted verbal contextual graph to an unweighted graph by highlighting the important edges and deemphasizing the trivial edges; (ii) we obtain a dominating set from the converted (unweighted) verbal contextual graph, which takes into account both essentiality and integrality. We detail these two sub-processes in the following two subsections.

3.3. Iterative edge weight adjustment

In the *verbal contextual graph*, there are often edges with small weight between two nodes (terms) that may be irrelevant; e.g., the edge between vertices t_2 and t_3 has a small weight of 0.2 in Fig. 1. A pair of terms often has a low weight for Lin's similarity and other measurements [6]. To find the core vertices, it is necessary to highlight the important edges and deemphasize (or filter out) trivial edges in a *verbal contextual graph*. We assume that the important edge in a *verbal contextual graph* is important to both its ends. In other words, the important edge should have a weight higher than the weights of other edges that are connected to the two ends (vertices) of the important edge. According to this assumption, we adjust the edge weight of a vertex iteratively according to the initial weight ratios:

$$w_{k+1}(e_{ab})=\frac{w_k(e_{ab})+\frac{w_0(e_{ab})}{\sum_{\forall x}w_0(e_{ax})}\cdot u(t_a)+\frac{w_0(e_{ab})}{\sum_{\forall y}w_0(e_{yb})}\cdot u(t_b)}{\sum_{\forall e\in E_i}w_0(e)+\sum_{\forall t\in V_i}u(t)}, \tag{6}$$

where $w_k(e_{ab})$ is the weight of edge e_{ab} in the k-th iteration, $\sum_{\forall x}w_0(e_{ax})$ is the sum of initial weights for all edges connected to vertex t_a, $u(t_a)$ is the weight value of tag vertex t_a, $\sum_{\forall e\in E_i}w_0(e)$ is the sum of initial weights of all edges and $\sum_{\forall t\in V_i}u(t)$ is the sum of weights of all vertices.

Taking the edge e_{23} in Fig. 1 as an example, $w_2(e_{23})$ is calculated as

$$\begin{aligned}
w_1(e_{23})
&=\frac{w_0(e_{23})+\frac{w_0(e_{23})}{\sum_{\forall x}w_0(e_{2x})}\cdot u(t_2)+\frac{w_0(e_{23})}{\sum_{\forall 2}w_0(e_{y3})}\cdot u(t_3)}{\sum_{\forall e\in E_i}w_0(e)+\sum_{\forall t\in V_i}u(t)}\\
&=\frac{0.2+\frac{0.2\times u(t_2)}{w_0(e_{21})+w_0(e_{23})+w_0(e_{24})}+\frac{0.2\times u(t_3)}{w_0(e_{13})+w_0(e_{23})+w_0(e_{43})}}{\sum_{\forall e\in E_i}w_0(e)+(u(t_1)+u(t_2)+u(t_3)+u(t_4))}\\
&=\frac{0.2+\frac{0.2\times 0.3}{0.8+0.2+0.6}+\frac{0.2\times 0.2}{0.4+0.2+0.9}}{0.8+0.5+0.4+0.2+0.6+0.9+1}\approx 0.06,
\end{aligned} \tag{7}$$

where $\sum_{\forall x}w_0(e_{2x})$ is equivalent to $w_0(e_{21})+w_0(e_{23})+w_0(e_{24})$ as edges e_{21}, e_{23} and e_{24} are connected to vertex t_2, and $\sum_{\forall y}w_0(e_{y3})$ is similarly equivalent to $w_0(e_{13})+w_0(e_{23})+w_0(e_{43})$. Note that the update sequence of the weight does not affect the final output of the weight, as the above update method only relies on initial weights.

A crucial issue is whether the proposed weight adjusting method is convergent or not in a finite number of iterations. The proof of convergence is given below.

Lemma 1. *$\forall e_{ab}\in E_i$, $w_k(e_{ab})$ converges.*

Proof. Proving the convergence of $w_k(e_{ab})$ is equivalent to proving the two following properties of $w_k(e_{ab})$.

1. $w_k(e_{ab})$ either increases or decreases monotonically and
2. $w_k(e_{ab})$ has an upper or lower bound respectively.

$$\begin{aligned}
&w_{k+1}(e_{ab})-w_k(e_{ab})\\
&\quad=\frac{w_k(e_{ab})+\frac{w_0(e_{ab})}{\sum_{\forall x}w_0(e_{ax})}\cdot u(t_a)+\frac{w_0(e_{ab})}{\sum_{\forall y}w_0(e_{yb})}\cdot u(t_b)}{\sum_{\forall e\in E_i}w_0(e)+\sum_{\forall t\in V_i}u(t)}-w_k(e_{ab}).
\end{aligned} \tag{8}$$

Since $u(t)$ is the normalized weight, its summation is a value of 1; i.e., $\sum_{\forall t\in V_i}u(t)=1$. $\sum_{\forall e\in E_i}w_0(e)$ and $\frac{w_0(e_{ab})}{\sum_{\forall x}w_0(e_{ax})}\cdot u(t_a)+\frac{w_0(e_{ab})}{\sum_{\forall y}w_0(e_{yb})}\cdot u(t_b)$ are two constants that do not change among iterations. For the sake of simple notation, we respectively denote the sum of $\sum_{\forall e\in E_i}w_0(e)$ and $\frac{w_0(e_{ab})}{\sum_{\forall x}w_0(e_{ax})}\cdot u(t_a)+\frac{w_0(e_{ab})}{\sum_{\forall y}w_0(e_{yb})}\cdot u(t_b)$ as S_0 and U_{ab}. The above equation then becomes

$$\begin{aligned}
&w_{k+1}(e_{ab})-w_k(e_{ab})\\
&\quad=\frac{w_k(e_{ab})+U_{ab}}{1+S_0}-w_k(e_{ab})\\
&\quad=\frac{w_k(e_{ab})+U_{ab}-w_k(e_{ab})-S_0\cdot w_k(e_{ab})}{1+S_0}\\
&\quad=\frac{\frac{U_{ab}}{S_0}-w_k(e_{ab})}{1+\frac{1}{S_0}}.
\end{aligned} \tag{9}$$

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

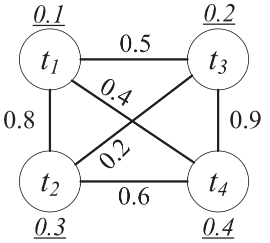


Fig. 1. Example of a verbal contextual graph.

For specific edge e_{ab}, the comparison between $\frac{U_{ab}}{S_0}$ and $w_k(e_{ab})$ is only related to the comparison between $\frac{U_{ab}}{S_0}$ and $w_0(e_{ab})$. When the edge obtains more (less) from the weight than the average gain of the other edges, its weight monotonically increases (decreases) until $\frac{U_{ab}}{S_0} = w_k(e_{ab})$. From Eq. (9), three scenarios can be summarized:

1. if $\frac{U_{ab}}{S_0} - w_0(e_{ab}) > 0$, $w_k(e_{ab})$ is a monotonically increasing sequence until $\frac{U_{ab}}{S_0} - w_k(e_{ab}) = 0$,
2. if $\frac{U_{ab}}{S_0} - w_0(e_{ab}) < 0$, $w_k(e_{ab})$ is a monotonically decreasing sequence until $\frac{U_{ab}}{S_0} - w_k(e_{ab}) = 0$,
3. if $\frac{U_{ab}}{S_0} - w_0(e_{ab}) = 0$, $w_k(e_{ab})$ is a constant sequence in equilibrium.

Additionally, since

$$w_{k+1}(e_{ab}) = \frac{w_k(e_{ab}) + U_{ab}}{1 + S_0}, \tag{10}$$

where $\frac{w_k(e_{ab}) + U_{ab}}{1+S_0}$ obviously has the upper bound $\frac{1+U_{ab}}{1+S_0}$ in the monotonically increasing case and lower bound $\frac{U_{ab}}{1+S_0}$ in the monotonically decreasing case, $\forall e_{ab} \in E_i$, $w_k(e_{ab})$ converges.□

The weights of edges converge to specific values when using the above weight adjusting method. We remove those edges with weights less than the first quartile (Q_1) when the weights are stable ($|w_{k+1}(e_{ab}) - w_k(e_{ab})| < 0.00001$, which can be ensured with 100 iterations). We keep the remaining edges and consider them important edges. As discussed above, we transform the weighted graph to an unweighted graph so that we can retain integrity. However, there is a special case ($1 = A_k + w_k(e_{ab})$) that the weight of an edge remains constant; e.g., if all edge and vertex weights in Fig. 1 are constants. In practice, it is almost impossible that all edge weights are equal as measured by Lin's similarity (as defined in (5)). It also rarely happens that all edge vertices are equal. To handle this extreme case, we can treat all edge weights as 1, since they are equally important. The *verbal contextual graph* is thus converted into an *unweighted verbal contextual graph*. The edges with zero weights are eliminated from the edge set. We denote the *unweighted verbal contextual graph* as $G'_i = \{V'_i, E'_i\}$ to differentiate it from the original graph. Note that the iterative update can be achieved with a closed-form update, as Eq. (6) is in the form of $\frac{w_k(e_{ab}) + U_{ab}}{1+S_0}$ (where $p = U_{ab}, q = 1 + S_0$), which can be transformed to $\frac{w_0(e_{ab}) + p(1 + q + \cdots + q^{k-1})}{q^k}$. The closed-form update can be more efficient in practice, while the iterative update illustrates the idea more clearly from the contextual graph perspective.

3.4. Graph dominating set discovery

In graph theory, the *dominating set* for a graph is a subset of the vertex set such that each vertex not in the *dominating set* is adjacent to at least one member of the set [15]. In Fig. 2, the blue vertices are dominating sets for the same graph in cases (a), (b) and (c). Different from methods that only focus on importance (e.g., PageRank and HITS), the dominating set provides a good balance essentiality and integrality because the dominating property (that each vertex not in the *dominating set* is adjacent to at least one of its members) ensures that the dominating vertices are not only more important than the remaining vertices (property of essentiality) but also represent the main characteristics of the graph (property of integrality).

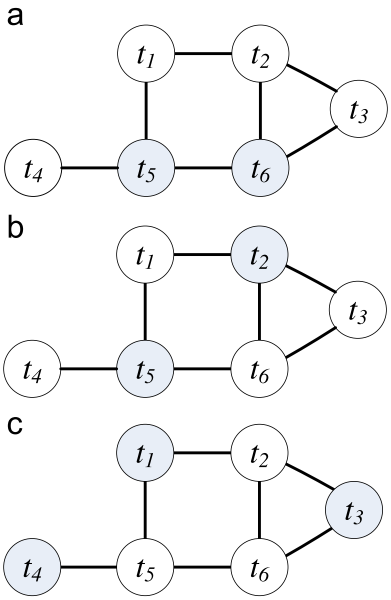


Fig. 2. Examples of the dominating set. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

Moreover, clustering algorithms (e.g., k-means and expectation–maximization algorithms) are unsuitable for obtaining the core set of vertices because (i) cluster-based methods identify several clusters whereas our goal is to find out a single core set and (ii) clustering processes need to tune variables, while the use of the dominating set is a non-parametric method that has a low computation load. The dominating set for the unweighted *verbal contextual graph* is formally defined below.

Definition 6. A *dominating set* for the *unweighted verbal contextual graph* G'_i, denoted by D_i, has two properties:

1. $D_i \subseteq V'_i$;
2. $\forall t_a \in V'_i - D_i$, $\exists t_x \in D_i$ and $e_{xa} \in E'_i$.

The first property states that D_i is a subset of vertex set V'_i, while the second property states that each vertex not in D_i is adjacent to at least one member of D_i.

Lemma 2. *$\forall G'_i$ satisfies $E'_i \neq \varnothing$, $\exists D_i, s.t. |D_i| < |V'_i|$.*

Proof. Since $V'_i \neq \varnothing$, let e_{ab} denote an edge in E'_i; i.e., $e_{ab} \in E'_i$. It is obvious that the dominating set D_i always exists as

$$D_i = V'_i - \{t_a\} \quad (\text{or } V'_i - \{t_b\}), \tag{11}$$

because D_i satisfies the two properties in Definition 6:

1. $D_i = V'_i - \{t_a\} \subseteq V'_i$
2. $\forall t_a \in V'_i - D_i$, which is equivalent to $V'_i - (V'_i - \{t_a\}) = \{t_a\}$, $\exists t_b \in V'_i - \{t_a\}$ and $e_{ab} \in E'_i$.

Additionally, since $|D_i| = |V'_i - \{t_a\}| = |V'_i| - 1$, $|D_i| < |V'_i|$. Thus, $\forall G'_i$ satisfies $E'_i \neq \varnothing$, $\exists D_i$, s.t. $|D_i| < |V'_i|$.□

According to Lemma 2, we can always find a *dominating set* smaller than vertex set in the *unweighted verbal contextual graph*. It is an intuitive step to discover the *dominating set* with minimal size (i.e., the minimum dominating set) that represents the *unweighted verbal contextual graph* (e.g., minimum dominating sets have size of 2 in cases (a) and (b) in Fig. 2). However, it has been proven that finding a minimum dominating set for a graph with n vertices is an NP-complete decision problem [12], and the state-of-the-art can discover a minimum dominating set with running time $O^*(2^{0.417n})$ under polynomial space [4]. Since this step must be achieved on-line, the efficiency is an essential factor to be taken into consideration. Therefore, we adapt the existing method [26] to this problem, and propose the following *Dominating Set Discovery Algorithm*.

Algorithm 1.

```
Data: the edge set $E'_i$ and the vertex set $V'_i$ ($|V'_i| = n$)
Result: a dominating set $D_i$
for each $t_x \in V'_i$ do
    if $deg(t_x) > 0$ then
        for each $e_{xy} \in E'_i$ do
            $V'_i \leftarrow V'_i - \{t_y\}$;
            $E'_i \leftarrow E'_i - \{e_{xy}\}$;
        end
    end
    $D_i \leftarrow D_i \cup \{t_x\}$;
end
```

As shown in Algorithm 1, the purpose of the *Dominating Set Discovery Algorithm* is to find a dominating set with acceptable time complexity rather than to find a minimum dominating set. The complexity of this algorithm is $O(\lambda \cdot n)$, where λ is the maximal degree of the graph. Therefore, the maximal degree of the graph should be significantly less than the number of vertices ($\lambda \ll n$) in most cases. By obtaining the *dominating set* from the *unweighted verbal contextual graph*, we define the *dominating context* to represent the pruned context based on the *dominating set* as follows.

Definition 7. Let $\{t_1^i, \dots, t_k^i\} = D_i$ and $\{\varepsilon_1^i, \dots, \varepsilon_k^i\}$ be terms in a dominating set and their contextual relevance to query q_i respectively. The *dominating context* is represented by the vector $\overrightarrow{c_i}$ as

$$\overrightarrow{c_i} = (t_1^i : \varepsilon_1^i, t_2^i : \varepsilon_2^i, \dots, t_k^i : \varepsilon_k^i).$$

The degree of relevance ε_a^i is obtained by the normalized frequency of the term appearing in the previous queries within a task session:

$$\varepsilon_a^i = \frac{f(t_a^i)}{\sum_{\forall j} f(t_j^i)}, \tag{12}$$

where $f(t_a^i)$ is the frequency of term t_a appearing in the previous queries and $\sum_{\forall j} f(t_j^i)$ is the sum of frequencies for all terms in the *dominating set*. A higher value of ε_a^i indicates that t_a^i is more important in the current context.

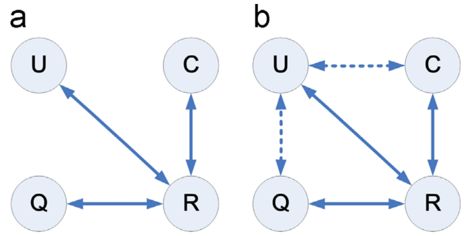


Fig. 3. Comparison of (a) cosine and (b) revised ranking models.

3.5. Resource ranking score models

3.5.1. Cosine ranking model

The elements in R, U, Q and C in Eq. (1) are defined by the resource profile $\overrightarrow{r_x}$, user profile $\overrightarrow{u_i}$, query $\overrightarrow{q_i}$ and dominating context $\overrightarrow{c_i}$ respectively. Intuitively, we can employ the cosine similarity measurement for resource ranking (as in other folksonomy-based applications [33,27]); i.e.,

$$\theta(\overrightarrow{r_x}, \overrightarrow{u_i}, \overrightarrow{q_i}, \overrightarrow{c_i}) = \frac{\overrightarrow{r_x} \cdot \overrightarrow{u_i}}{\| \overrightarrow{r_x} \| | \overrightarrow{u_i} \|} \cdot \frac{\overrightarrow{r_x} \cdot \overrightarrow{q_i}}{\| \overrightarrow{r_x} \| | \overrightarrow{q_i} \|} \cdot \frac{\overrightarrow{r_x} \cdot \overrightarrow{c_i}}{\| \overrightarrow{r_x} \| | \overrightarrow{c_i} \|}. \tag{13}$$

A higher value of θ indicates that the resource is more suitable in terms of the query, user preference and context. The personalized resource ranking is based on this θ function. Note that the paradigm of the cosine ranking method is to measure the relevance between each resource and context (query or user profile) separately, and then aggregate the relevance scores into a single score as shown in Fig. 3(a).

However, the cosine ranking model suffers the problem of the uniform treatment of tags in the user profile [32]. Specifically, the cosine ranking paradigm neglects the relationships between queries (contexts) and user profiles. For example, a user profile contains two tags “icecream” and “spicy”, and only the tag “spicy” in the user profile may be relevant when the user issues the query “braised beef”. To prevent this from being problematic, we propose the following revised ranking model.

3.5.2. Revised ranking model

As illustrated by the dashed arrows in Fig. 3(b), we further incorporate the relationships between queries (contexts) and user profiles in the revised ranking model to tackle the aforementioned problem of the uniform treatment of tags in the user profile. According to our observation, a tag is relevant to the current query (context) if it co-occurs with any terms in the current query (context) that annotate the same resource. For example, if a user profile contains two tags “ice-cream” and “spicy”, the tag “spicy” is the term relevant to the current query “braised beef” as it should co-occur with query terms “braised” or “beef” that annotate a resource like “fried steak”, while the tag “icecream” is irrelevant as it is normally not used to describe the same resources as “braised” or “beef”. Therefore, the following piecewise function is proposed to distinguish relevant tags in the user profile under the current query context:

$$\tau_n^{*i} = \begin{cases} \tau_n^i & \exists a, \exists x \text{ s.t. } \{t_n^i, t_a\} \subseteq \overrightarrow{r_x} \ (t_n^i \in \overrightarrow{u_i}, t_a \in \overrightarrow{q_i} \cup \overrightarrow{c_i}), \\ 0 & \text{otherwise}, \end{cases} \tag{14}$$

where t_n^i and τ_n^i are the tag and the corresponding relevance degree in the user profile (as defined Definition 2), and t_a is a tag in query ($\overrightarrow{q_i}$) or context ($\overrightarrow{c_i}$).

Since all relevance degrees in user profile $\overrightarrow{u_i}$ change from τ_n^i to τ_n^{*i}, we debite the updated user profile as $\overrightarrow{u'_i}$, and then adopt the cosine similarity measurement, which is the same as the cosine ranking model:

$$\theta'(\overrightarrow{r_x}, \overrightarrow{u'_i}, \overrightarrow{q_i}, \overrightarrow{c_i}) = \frac{\overrightarrow{r_x} \cdot \overrightarrow{u'_i}}{\| \overrightarrow{r_x} \| | \overrightarrow{u'_i} \|} \cdot \frac{\overrightarrow{r_x} \cdot \overrightarrow{q_i}}{\| \overrightarrow{r_x} \| | \overrightarrow{q_i} \|} \cdot \frac{\overrightarrow{r_x} \cdot \overrightarrow{c_i}}{\| \overrightarrow{r_x} \| | \overrightarrow{c_i} \|}. \tag{15}$$

Note that the only difference between θ' and θ is the updated (or contextualized) user profile $\overrightarrow{u'_i}$ whose relevance degrees are updated using (14).

A more flexible approach θ^* allocates weights to each component in θ' (or θ). In this case, the weighted sum is the ranking score:

$$\theta^*(\vec{r_x},\vec{u_i},\vec{q_i},\vec{c_i}) = \alpha\cdot\frac{\vec{r_x}\cdot\vec{u_i}}{\|\vec{r_x}\|\,|\,\vec{u_i}\|} + \beta\cdot\frac{\vec{r_x}\cdot\vec{q_i}}{\|\vec{r_x}\|\,|\,\vec{q_i}\|} + \gamma\cdot\frac{\vec{r_x}\cdot\vec{c_i}}{\|\vec{r_x}\|\,|\,\vec{c_i}\|}, \quad (16)$$

where α, β and γ are parameters that control the effects of the three components and satisfy $\alpha,\beta,\gamma\in[0,1]$ and $\alpha+\beta+\gamma=1$. We will investigate the effects of these parameters in the following experiment.

4. Experiment and findings

To verify the effectiveness of our proposed method, we conduct experiments using the Movielens dataset. We compare our method with baselines in terms of personalized search performance.

4.1. Experiment setup

4.1.1. Dataset

Movielens has 10,681 resources (movies), 10,000,054 tags and 71,567 users. The tags depict various aspects from the movie intrinsic content to user extrinsic perception. Employing the criteria of the inactive user described in [24], we select users who have tagged no less than 15 resources and exclude inactive users.

The format of the dataset is a quadruplet, which includes the user, tags (query), resource and time stamp. The dataset is randomly split into a training set (80%) and test set (20%) for each user. For each test tuple, the resource selected by the user is considered as the ground truth. We examine the accuracy and effectiveness of the proposed approach and baselines in predicting the target resource by giving the query terms (tags) for each user in the test set.

4.1.2. Metric

Three widely used metrics are employed in our experiments, namely *P@N* (Precision @N) [30], *MRR* (Mean Reciprocal Rank) and *RRI* (Related Ranking Improvement) [25]. The metric *P@N*, measuring the accuracy of the personalized search strategy, is defined as

$$P@N = \sum_{i=1}^{n}\frac{p(q_i)}{n}, \quad (17)$$

$$p(q_i) = \begin{cases} 1 & \text{if } rank(r_i^q) \le N, \\ 0 & \text{if } rank(r_i^q) > N, \end{cases} \quad (18)$$

where $rank(R_i^q)$ is the rank of the target resource for query q_i and n is the number of queries in the test. The metric *MRR*, which denotes the mean rank of target resources, quantifies how quickly a personalized strategy can assist users in finding relevant resources. The metric *RRI* is derived from *MRR* and represents the improvement in performance of the personalized search method over the baseline method. These two metrics are defined as

$$MRR = \frac{1}{n}\cdot\sum_{i=1}^{n}\frac{1}{rank(r_i^q)}, \quad (19)$$

$$RRI = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(\frac{rank_b(r_i^q)}{rank_a(r_i^q)} - 1\right) = \frac{MRR_a - MRR_b}{MRR_b}, \quad (20)$$

where $rank_a(r_i^q)$ and $rank_b(r_i^q)$ are the ranks of the target resource obtained by two methods a and b, respectively.

4.1.3. Baseline

There are four baselines to be compared with our proposed method, and some are adapted from state-of-art methods. For simplicity of notation, we refer to our proposed method as “*DomContext*”. The abbreviations and details of the baselines are introduced below.

Basic: The basic method was proposed in [33] and depends only on the matching between the query (Q) and resource profile (R). In the experiment, we adopt the same parameter settings and ranking functions in [33], and the only difference is that we adopt the NTF (as discussed in [7]) as a more effective paradigm to replace the original TF-IUF/IRF in the construction of the user and resource profiles. The comparison of different approaches is unaffected by the use of different paradigms. Note that neither the user profile (U) or context (C) (1) is included in this baseline.

Uncontext: The second method does not include any context in the resource ranking. In other words, this baseline mainly relies on the query (Q), user (U) and resource profiles (R). Context (C) is excluded from Eq. (1) for this baseline, and the method is essentially the same as that in [27]. We again replace the original paradigm with the NTF and keep all other settings.

Context: The third method adopts the original context set rather than using the dominating set, and the degree of importance of terms in the context is normalized by the sum frequency of terms in the previous queries within a session. Note that the *Context* method is similar to the method in [32]. This baseline also uses the NTF, and no modifications are needed.

MinContext: The baseline always employs the minimum dominating set as the dominating context, which can be regarded as an extreme case for our proposed *DomContext* method. Note that the running time of this baseline increases exponentially ($O^*(2^{0.417n})$) as mentioned in Section 3.

4.2. Comparison of overall performance

For the methods described above, we adopt the NTF as the user and resource profiling paradigms, and employ the cosine ranking model to rank the resources. We will discuss other paradigms and the revised ranking model later. The performances of the proposed and baseline approaches in terms of *P@N* are illustrated in Fig. 4. The following findings are taken from the results.

(1) It is beneficial to incorporate the verbal context, since the context not only allows exploitation of the hidden intention of the query but also the eliminates ambiguity. The supporting evidence is that methods with context (*Context*, *MinContext* and

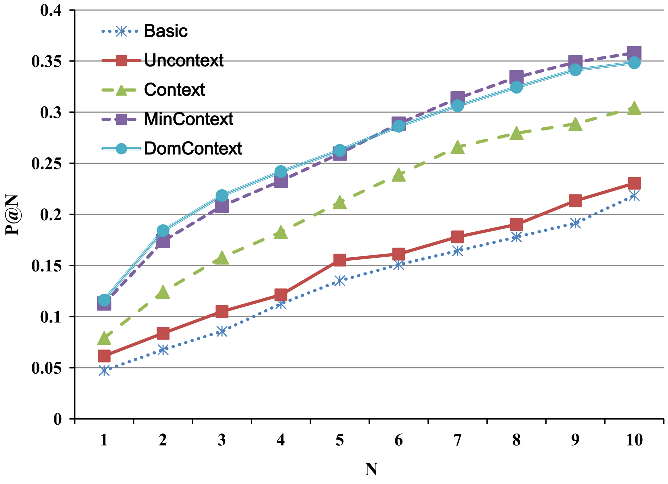


Fig. 4. P@N for the proposed and baseline methods.

DomContext) perform better than methods without context (*Uncontext* and *Basic*) in terms of *P@N*. To simplify the notion, we use the symbol " $>$ " to denote better performance:

$DomContext, MinContext, Context > Uncontext, Basic$

(2) It is better to obtain the dominating set from the verbal contextual graph and use the set instead of using the whole verbal context directly, because the later may include noisy contextual elements in the process of finding a relevant resource, while the former mainly highlights the essential elements and filters out redundant and irrelevant elements, which can give results that are more precise. The supporting evidence is that

$DomContext, MinContext > Context$

Both *MinContext* and *DomContext* obtain a dominating set from the verbal contextual graph, while *Context* uses the whole verbal context directly.

(3) It is adequate to find the dominating set rather than the minimum dominating set to facilitate the personalized search. Because *MinContext* is much slower than *DomContext* as discussed in Section 3, while their performances are not significantly different from each other ($p > 0.5$ in a sign test). The supporting evidence is

$DomContext \approx MinContext$

MinContext is not an improvement of *DomContext* because the performance of the minimum dominating set is negatively affected by the problem of data sparsity.

(4) Extracting user profiles obviously improves the performance of the personalized search. This observation is consistent with the results of our earlier work [7]. Moreover, *Uncontext* improves on *Basic* by incorporating the user profile during resource ranking. The supporting evidence is

$UnContext > Basic$

We also performed a sign test to verify this relation and $p < 0.01$ further validates the observation.

Performances in terms of *MRR* and *RRI* are summarized in Fig. 5 and Table 2, respectively. The experimental results clearly show that *DomContext* method has the largest MRR value (0.1903), which is 21.83% to 80.38% better than the values for *Basic*, *Uncontext* and *Context*. *DomContext* also achieves slightly (not significantly) better performance than *MinContext* (by 0.32%). Note that values in Table 2 are normally asymmetric owing to the definition of *RRI* with an asymmetric property, as expressed by Eq. (20). Moreover, we found that the performances on *MRR* and *RRI* have trends similar to those of performances in terms of *P@N*.

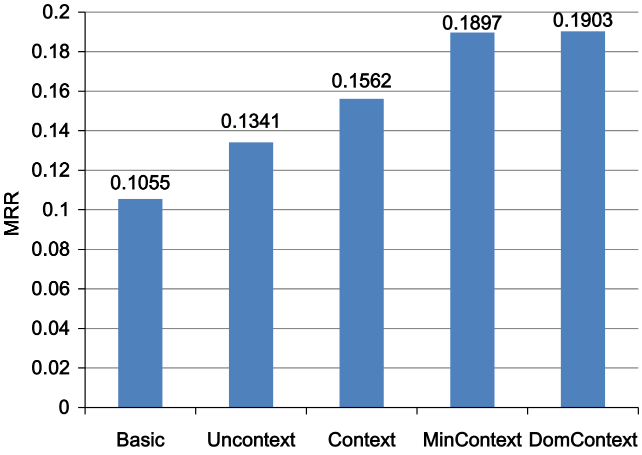


Fig. 5. MRR for the proposed and baseline methods.

4.3. Effects of various paradigms

As mentioned in Section 3., there are alternative paradigms such as TF, TF-IUF/IRF and BM 25 for the extraction of user and resource profiles. To illustrate their effects on the above context-aware methods (*Context, MinContext and DomContext*), we further compare the MRR for four different paradigms ($p < 0.03$ via the sign test) as shown in Fig. 6. It is clear that the NTF paradigm achieves the best performance among all paradigms whatever context-aware method is adopted, which is consistent with the conclusion that the NTF is the paradigm most suitable for user and resource profiling drawn in our previous work [7]. Moreover, TF has the worst performance since the absolute tag frequency of a tag does not indicate that the tag will be highly relevant to the user (or salient to the resource).

4.4. Effects of ranking models

We further compare two ranking models discussed in Section 3.5. The MRR values ($p < 0.01$ via the sign test) for the cosine ranking and revised ranking models used in three context-aware methods (*Context, MinContext and DomContext*) are shown in Fig. 7. The revised ranking model clearly performs better than the cosine ranking model, thus verifying our observation that not all tags in user profile are relevant to current queries (contexts). Note that there is no significant performance difference between *MinContext* and *DomContext* whatever paradigm or ranking model is adopted.

4.5. Effects of parameters

As shown by Eq. (16), the parameters for three different components may also affect the search performance. To investigate their effect, we employ a grid search method, which searches in a two-dimensional space; $\alpha, \beta \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, giving a total of 21 combinations when filtering out for which $\alpha + \beta > 1$. We do not include γ here as it is

Table 2
RRI performance.

	Basic	Uncontext	Context	MinContext	DomContext
Basic	0.00%	−21.33%	−32.46%	−44.38%	−44.56%
Uncontext	27.11%	0.00%	−14.15%	−29.31%	−29.53%
Context	48.06%	16.48%	0.00%	−17.65%	−17.92%
MinContext	79.81%	41.46%	21.45%	0.00%	−0.32%
DomContext	80.38%	41.91%	21.83%	0.32%	0.00%

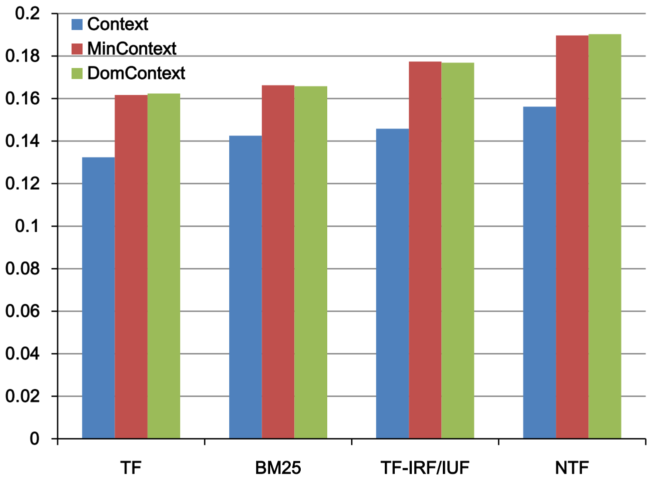


Fig. 6. MRR for four paradigms used in context-aware methods.

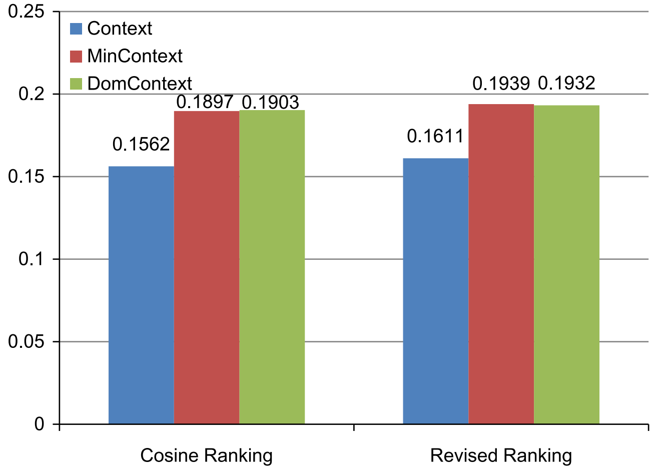


Fig. 7. Comparisons between cosine/revised ranking models.

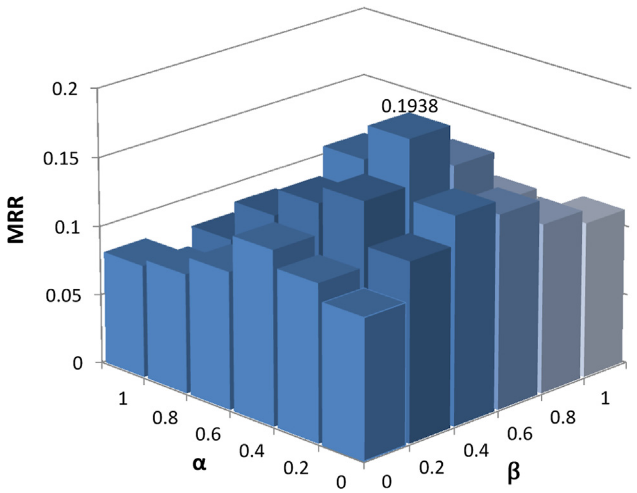


Fig. 8. MRR for different combinations of parameters.

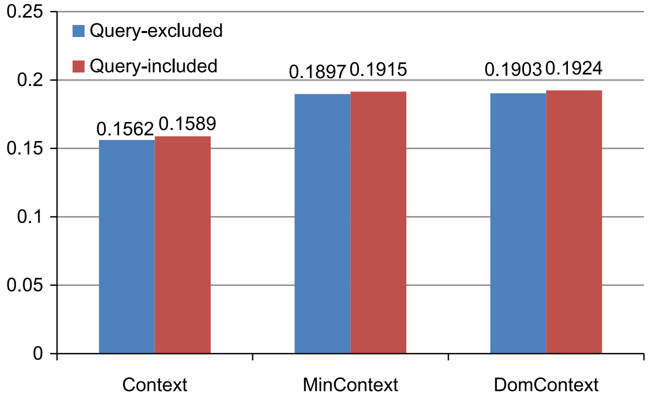


Fig. 9. Comparisons between query-included/excluded context graphs.

an independent variable ($\gamma = 1 - \alpha - \beta$). The *MRR* performances for varying parameters α and β are illustrated in Fig. 8. The combination having the best performance among all 21 combinations is $\alpha = 0.2$ and $\beta = 0.4$, with the *MRR* value being 0.1938, as the three components in Eq. (16) are reasonably weighted. Note that the absence of any component may hurt the performance seriously as there is an obvious gap in performance between involving three components and only involving two of them (e.g., from $\alpha = 0, \beta = 0.2$ to $\alpha = 0.2, \beta = 0.2$). In other words, these three components are important in realizing optimal performance. As the optimal combination of parameters may be different when the proposed model is applied in a different domain, we can adopt Eq. (15) for cross-domain applications. If we need to optimize the performance in a domain-specific application, we can further split 20% from the training set (80% of the dataset) as the validation set (16% of the dataset) and perform five-fold cross validation to tune the parameters.

4.6. *Effects of context graphs*

As mentioned in Definition 5, there are two types of context graphs, namely query-included and query-excluded context graphs. The distinction between the two types is whether query terms (tags) of the current query are included in the construction of the context graph. In this subsection, we compare the effects of the different types on three context-aware methods (i.e., *Context, MinContext and DomContext*). Fig. 9 shows that all context-aware methods employing a query-included context graph outperform methods employing a query-excluded context graph in terms of *MRR* value ($p < 0.01$ via a sign test). A reasonable explanation for this experimental result is that terms (tags) of the current query reflect user intention more precisely. If query terms (tags) of the current query are taken into consideration, a more precise contextual graph can be constructed so that the performance in terms of *MRR* can be improved.

5. Conclusion

This paper addressed (i) the construction of a verbal contextual graph to describe search contexts in folksonomy, (ii) the identification of core contextual elements and de-emphasis of trivial elements in verbal contexts, and (iii) the facilitation of a personalized search using different ranking models in folksonomy. To this end, we built a verbal contextual graph by connecting elements (terms) according to their semantic similarity measurement. Furthermore, the iterative weight adjustment method, which is proven to be convergent in a few iterations, transforms a verbal contextual graph to an unweighted one. According to two properties (essentiality and integrality), we argued that the dominating set in graph theory is a good choice and proposed an algorithm that obtains the dominating set with reasonable time complexity $O(\lambda \cdot k)$. To validate the effectiveness of our proposed method, we conducted experiments on a Movielens dataset by comparing the method with baselines in terms of personalized search performance. The experimental result verified our observations and demonstrated that our proposed dominating method outperforms state-of-the-art baselines in terms of the personalized resource search. In our future research, we plan to continue studying other structures of user and resource profiles such as those having high-order graphs.

Acknowledgment

The work described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS11/E06/14).

References

[1] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, Alexander Tuzhilin, Incorporating contextual information in recommender systems using a multidimensional approach, ACM Trans. Inf. Syst. 23 (1) (2005) 103–145.

[2] Gediminas Adomavicius, Alexander Tuzhilin, Context-aware recommender systems, in: Recommender Systems Handbook, Springer US, New York, NY, USA, 2011, pp. 217–253.

[3] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, Zhong Su, Optimizing web search using social annotations, in: Proceedings of the 16th International Conference on World Wide Web, ACM, New York, NY, USA, 2007, pp. 501–510.

[4] Nicolas Bourgeois, F. Della Croce, B. Escoffier, V. Th Paschos, Fast algorithms for min independent dominating set, Discrete Appl. Math. 161 (4) (2013) 558–572.

[5] Sergey Brin, Lawrence Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1) (1998) 107–117.

[6] A. Budanitsky, G. Hirst, Semantic distance in wordnet: an experimental, application-oriented evaluation of five measures, in: Workshop on WordNet and Other Lexical Resources, vol. 2, NAACL, 2001.

[7] Yi Cai, Qing Li, Personalized search by tag-based user profile and resource profile in collaborative tagging systems, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2010, pp. 969–978.

[8] Iván Cantador, Pablo Castells, Semantic contextualisation in a news recommender system, in: Proceedings of the 1st International Workshop on Context-Aware Recommender Systems, ACM, New York, NY, USA, 2009.

[9] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, Qiang Yang, Context-aware query classification, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2009, pp. 3–10.

[10] D. Carmel, H. Roitman, E. Yom-Tov, Social bookmark weighting for search and recommendation, VLDB J. 19 (6) (2010) 761–775.

[11] Anita Fetzer, Recontextualizing Context: Grammaticality Meets Appropriateness, John Benjamins Publishing Company, Amsterdam, The Netherlands, 2004.

[12] Michael R. Garey, David S. Johnson, Computers and Intractability A Guide to the Theory of NP-Completeness, W. H. Freeman & Co., San Francisco, CA, USA, 1990.

[13] Mohsen Ghadessy, Text and Context in Functional Linguistics, vol. 169, John Benjamins Publishing Co, Amsterdam, The Netherlands, 1999.

[14] Isabelle Guyon, André Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[15] S.T. Hedetniemi, R.C. Laskar, Bibliography on domination in graphs and some basic definitions of domination parameters, Discrete Math. 86 (1–3) (1991) 257–277.

[16] Ting Jin, Haoran Xie, Jingsheng Lei, Qing Li, Xiaodong Li, Xudong Mao, Yanghui Rao, Finding dominating set from verbal contextual graph for personalized search in folksonomy, in: Proceedings of the 2013 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE, Washington, DC, USA, 2013, pp. 347–352.

[17] Jon M Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (5) (1999) 604–632.

[18] Andrej Košir, Ante Odic, Matevz Kunaver, Marko Tkalcic, Jurij F Tasic, Database for contextual personalization, Elektrotehn. Vestn. 78 (5) (2011) 270–274.

[19] Zhen Liao, Daxin Jiang, Enhong Chen, Jian Pei, Huanhuan Cao, Hang Li, Mining concept sequences from large-scale search logs for context-aware query suggestion, ACM Trans. Intell. Syst. Technol. 3 (1) (2011) 17.

[20] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the 15th International Conference on Machine Learning, IMLS, vol. 1, 1998, pp. 296–304.

[21] Marcelo G Manzato, Rudinei Goularte, A multimedia recommender system based on enriched user profiles, in: Proceedings of the 27th Annual ACM Symposium on Applied Computing, ACM, New York, NY, USA, 2012, pp. 975–980.

[22] Michael G Noll, Christoph Meinel, Web search personalization via social bookmarking and tagging, in: Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 367–380.

[23] Umberto Panniello, Alexander Tuzhilin, Michele Gorgoglione, Cosimo Palmisano, Anto Pedone, Experimental comparison of pre-vs. post-filtering approaches in context-aware recommender systems, in: Proceedings of the Third ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2009, pp. 265–268.

[24] Jitao Sang, Xu Changsheng, Liu. Jing, User-aware image tag refinement via ternary semantic analysis, IEEE Trans. Multimedia 14 (3) (2012) 883–895.

[25] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, Robin Burke, Personalized recommendation in social tagging systems using hierarchical clustering, in: Proceedings of the 2008 ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2008, p. 259–266.

[26] Kuo-Hui Tsai, Wen-Lian Hsu, Fast algorithms for the dominating set problem on permutation graphs, in: Algorithms, Springer, Berlin Heidelberg, 1990, pp. 109–117.

[27] David Vallet, Iván Cantador, Joemon M Jose, Personalizing web search with folksonomy-based user and document profiles, Advances in Information Retrieval, Springer, Berlin, Heidelberg (2010) 420–431.

[28] Vargas-Govea Blanca, González-Serna Gabriel, Ponce-Medellin Rafael, Effects of Relevant Contextual Features in the Performance of a Restaurant Recommender System. Context Aware Recommender Systems, ACM, New York, NY, USA, 2011.

[29] Xinxi Wang, David Rosenblum, Ye Wang, Context-aware mobile music recommendation for daily activities, in: Proceedings of the 20th ACM International Conference on Multimedia, ACM, New York, NY, USA, 2012, pp. 99–108.

[30] Ryen W White, Peter Bailey, Liwei Chen, Predicting user interests from contextual information, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2009, pp. 363–370.

[31] Ryen W White, Steven M Drucker, Investigating behavioral variability in web search, in: Proceedings of the 16th International Conference on World Wide Web, ACM, New York, NY, USA, 2007, pp. 21–30.

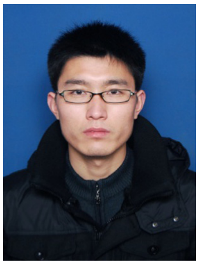
[32] Haoran Xie, Qing Li, Xudong Mao, Context-aware personalized search based on user and resource profiles in folksonomies, in: Web Technologies and Applications, Springer, Berlin, Heidelberg, 2012, pp. 97–108.

[33] Shengliang Xu, Shenghua Bao, Ben Fei, Zhong Su, Yong Yu, Exploring folksonomy for personalized search, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2008, pp. 155–162.

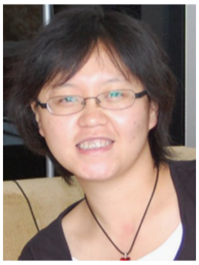
[34] Yong Zheng, Robin Burke, Bamshad Mobasher, Differential context relaxation for context-aware travel recommendation, E-Commerce and Web Technologies, Springer, Berlin, Heidelberg (2012) 88–99.

Haoran Xie is an assistant professor in Caritas Institute of Higher Education. He received his Ph.D. and M.Sc. from Department of Computer Science, City University of Hong Kong, and Bachelor of Engineering from School of Software Engineering, Beijing University of Technology. Before he joined CIHE, he was a senior research assistant in Hong Kong Baptist University. His research interests include social media, big data, data mining, recommender systems, human computer interaction and e-learning systems.

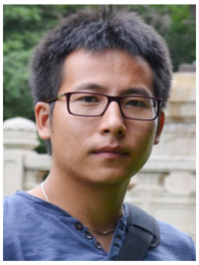
Xiaodong Li was born in Nanjing, China, in 1984. He received the B.Sc. in computer science and technology from the Nanjing University, China, in 2007. He is currently pursuing the Ph.D. degree at City University of Hong Kong. His current research interests include the machine learning, data mining and market micro structure.



Tao Wang was born in Jiangxi, China, in 1990. He is currently a M.Sc. student of School of Software Engineering at the South China University of Technology, Guangzhou, China. His current research interests include opinion mining, text mining algorithms and statistical machine learning techniques.



Li Chen is an assistant professor at Hong Kong Baptist University. She obtained her Ph.D. degree in human computer interaction at Swiss Federal Institute of Technology in Lausanne (EPFL), and Bachelor and Master degrees in computer science at Peking University, China. Her research interests are mainly in the areas of human–computer interaction, user-centered development of recommender systems and e-commerce decision supports. Her co-authored papers have been published in journals and conferences on e-commerce, artificial intelligence, intelligent user interfaces, user modeling, and recommender systems.



Ke Li was born in Hunan, China, in 1985. He received the B.Sc. and M.Sc. degrees in computer science and technology from the Xiangtan University, China, in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree at City University of Hong Kong. His current research interests include the evolutionary multi-objective optimization, surrogate assisted evolutionary algorithms and statistical machine learning techniques.

Fu Lee Wang is a professor and vice-president (Research and Advancement) in Caritas Institute of Higher Education. He received B.Eng. and M.Phil. from The University of Hong Kong, M.Sc. from The Hong Kong University of Science and Technology, MBA from Imperial College London, and Ph.D. from The Chinese University of Hong Kong. His research interests include electronic business, information retrieval, information systems and e-learning. Before joining Caritas, he was a faculty member at the City University of Hong Kong. He is Past Chair of ACM Hong Kong Chapter and Chair of IEEE Hong Kong Section Computer Society Chapter. He served as programme/conference chair of a number of international conferences.

Yi Cai received the Ph.D. degree in computer science from The Chinese University of Hong Kong. He is currently an associate professor of School of Software Engineering at the South China University of Technology, Guangzhou, China. His research interests are recommendation system, personalized search, semantic web and data mining.

Qing Li is a professor at the City University of Hong Kong. His research interests include object modeling, multimedia databases, social media and recommender systems. He is a Fellow of IET, a senior member of IEEE, a member of ACM SIGMOD and IEEE Technical Committee on Data Engineering. He is the chairperson of the Hong Kong Web Society, and is a steering committee member of DASFAA, ICWL, and WISE Society.

Huaqing Min is a professor and the dean of School of Software Engineering, South China University of Technology, China. His research interest includes artificial intelligence, machine learning, database, data mining and robotics.