

# Dynamic Context-Sensitive PageRank for Expertise Mining\*

Daniel Schall and Schahram Dustdar

Distributed Systems Group, Vienna University of Technology  
Argentinerstr. 8/184-1, A-1040 Vienna, Austria  
{schall,dustdar}@infosys.tuwien.ac.at

**Abstract.** Online tools for collaboration and social platforms have become omnipresent in Web-based environments. Interests and skills of people evolve over time depending in performed activities and joint collaborations. We believe that ranking models for recommending experts or collaboration partners should not only rely on profiles or skill information that need to be manually maintained and updated by the user. In this work we address the problem of expertise mining based on performed interactions between people. We argue that an expertise mining algorithm must consider a person's interest and activity level in a certain collaboration context. Our approach is based on the PageRank algorithm enhanced by techniques to incorporate contextual link information. An approach comprising two steps is presented. First, offline analysis of human interactions considering tagged interaction links and second composition of ranking scores based on preferences. We evaluate our approach using an email interaction network.

## 1 Introduction

The collaboration landscape has changed dramatically over the last years by enabling users to shape the Web and availability of information. While in the past collaborations were bounded to intra-organizational collaborations using a companies specific platform, and also limited to messaging tools such as email, it is nowadays possible to utilize the knowledge of an immense number of people participating in collaborations on the Web. The shift toward the Web 2.0 allows people to write blogs about their activities, share knowledge in forums, write Wiki pages, and utilize social platforms to stay in touch with other people.

Task-based platforms for human computation (i.e., *crowdsourcing*) including Amazon Mechanical Turk<sup>1</sup> enable access to the manpower of thousands of people on demand by creating human-tasks that are processed by the crowd. Human computation techniques applied in service-oriented environments involving Human-Provided Services (HPS) [1] enable the dynamic discovery and composition of experts. In such open and dynamic environments it becomes essential to manage expertise profiles and reputation of people in an automated manner.

---

\* This work is supported by the EU FP7 project COIN (No. ICT-2008-216256).

<sup>1</sup> Amazon Mechanical Turk: <http://www.mturk.com/>

Somebody seeking help or advice on a specific problem or businesses issuing task requests using, for example, before mentioned platforms need to be able to find the right person who can assist with offering his/her expertise. Recent work in expert finding (e.g., [2]) has been addressing the search for persons with the right skill level by using ontologies and by combining diverse semantic information from user/skill profiles. Since Web-scale collaborations involving a large amount of people not only demand for scalable algorithms and ranking solutions, but in many cases it is also desirable to consider the global properties of a human interaction network to determine the importance of users. Algorithms and models designed for ranking Web documents in the Web graph such as *PageRank* [3] or the *Hubs and Authorities* model (Hyperlink-Induced Topic Search [4] abbreviated as HITS) have been applied in online communities (e.g., Java discussion forums) [5,6] for expertise mining.

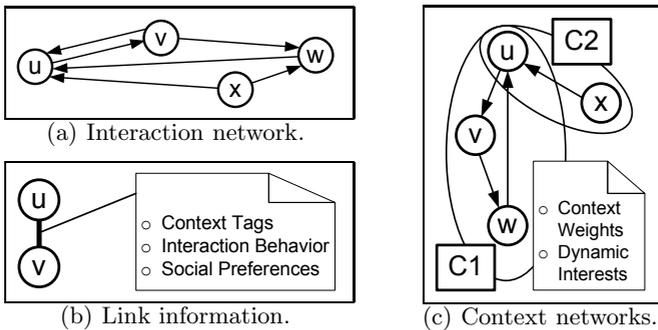
In this paper we present the following key contributions:

- We propose link analysis techniques derived from the popular PageRank model. Our approach is called *Dynamic Skill and Activity-based PageRank* (DSARank) accounting for activity and contextual link information.
- We present metrics to capture dynamic properties in human collaborations.
- Evaluation and discussion of our approach using a real email interaction dataset. Our experiments confirm that our proposed ranking approach is well suited for expertise mining.

## 2 Techniques for Expertise Ranking

### 2.1 Context-Sensitive Interaction Mining

Our approach is based on mining of interactions between users. Therefore, an interaction network (see Fig. 1(a)) is modeled as a graph  $G = (N, E)$  composed of the set of nodes  $N$  and the set of edges  $E$ . Note, in this work the terms edge and link have the same meaning.



**Fig. 1.** Collaborative networks: (a) interactions are performed between nodes in the network; (b) Metadata and metrics are associated with links between nodes; (c) Context networks are created based in link information

We argue that context information is essential for expertise mining. The context of an interaction can be captured by, for example, extracting relevant keywords from messages exchanged between users or by tags applied to various collaboration artifacts. In this work, we focus on *tags* (Fig. 1(b)) serving as input for contextual link information. Interaction metrics such as weights depicting the interest and focus of a user to collaborate with other peers in a specific context are automatically calculated through mining. Fig. 1(c) shows networks for context  $C1$  and  $C2$ . Each context network may have one or more tags associated with it.

Existing work in the area of expertise mining (e.g., [5,6]) typically focuses on a graph representation as depicted by Fig. 1(a). In contrast, we present an approach and algorithm that is suitable for scenarios as shown in Fig. 1(c). Before presenting the detailed mechanisms, we overview existing link analysis techniques.

## 2.2 Link Analysis Approaches

We focus on two well-established techniques (i.e., PageRank and HITS) that have attracted extensive research (see [7]) in the information retrieval and theoretical computer science community for over a decade. Here we outline the basic principles of both approaches and discuss in which cases one technique should be favored over the other in expertise mining applications. Using the well-known PageRank [3], the importance of a node  $u$  is defined as:

$$PR(u) = \overbrace{\frac{(1-\alpha)}{|N|}}^{\text{personalization}} + \alpha \cdot \overbrace{\sum_{(v,u) \in E} \frac{PR(v)}{\text{outdegree}(v)}}^{\text{importance propagation}} \quad (1)$$

The equation consists of two components. The first part (*personalization*) can be used to assign preferences to certain nodes. Without any preferences or bias towards a particular node, a personalization vector  $p(u) = 1/|N|$  is assumed. The second part of Eq. 1 propagates importance scores between connected nodes. Page and Brin [3] originally proposed a ‘teleportation’ factor  $\alpha = 0.85$  based on the *random surfer model*. The damping factor  $d = (1 - \alpha)$  is based on the assumption that a user on the Web typically follows six links ( $d = 1/6 = 0.15$ ).

A quite intuitive justification for applying PageRank in expertise ranking applications is the expected ‘informedness’ of users. Each link between people can be interpreted as a channel to propagate information in the network. The strength of a link limits the flow of information between two nodes  $v$  and  $u$ .

**Table 1.** PageRank and related symbols

Symbol	Description
$\alpha$	The PageRank ‘teleportation’ factor (a value between 0.8 and 0.9).
$ N $	The number of nodes in $G$ .
$PR$	The PageRank vector $PR$ holding importance scores for each node.
$\text{outdegree}(v)$	The number of links originating from $v$ .

For example,  $v$  may notify one of its neighbors  $u$  about some news or forward information to a newly chosen node.

The second approach discussed in this work for expertise ranking is HITS [4]. The HITS model has been proposed for expertise ranking in, for example, question and answer communities (Q/A) [6]. In such environments, hubs are characterized by users posting questions attracting answers from other knowledgeable users, representing authorities in the Q/A community.

$$\begin{aligned} H(u) &= \sum_{(u,v) \in E} A(v) \\ A(v) &= \sum_{(u,v) \in E} H(u) \end{aligned} \quad (2)$$

The benefit of the HITS approach is the availability of two ranking scores for each node (see Table 2). Using the concept of hubs and authorities is beneficial in task-based *online help and support* environments to support the coordination of delegation flows [8].

**Table 2.** HITS and related symbols

Symbol	Description
$H(u)$	Hub score of $u$ characterized by the authority of nodes $u$ points to.
$A(v)$	Authority score of $v$ influenced by hubs pointing to $v$ .

To briefly outline this idea. A member of the online help and support community (modeled as HPS [1]) may receive a request for support (RFS) and delegate work to some other peer in the network (characterizing hubs in the network). Receivers of the delegated work, however, expect RFSs fitting their skills and expertise (i.e., being an authority in the given domain). Careless delegations of work will overload these peers resulting in degraded processing time due to missing expertise. Within the online help and support community, authorities give feedback using rating mechanism (e.g., a number on the scale from 1 to 5) to indicate their satisfaction - whether a particular hub distributes work according to their skills and interest. A ‘good hub’ is thereby characterized by a neighborhood of peers that are satisfied with received RFSs. On the other hand, delegation of work is strongly influenced by trust, for example, whether the initial receiver of the RFS (hub within the community) expects that the peer will process work in a reliable and timely manner. Thus, RFS receivers need to be trusted by influential hubs that are highly rated in order to be recognized as authoritative peers in the community.

### 2.3 PageRank vs. HITS

Both PageRank and HITS follow similar principles by making use of the link structure in networks to estimate the importance of nodes. The benefit of

PageRank is that it takes the global network into account. The PageRank model is based on the interpretation of a graph  $G$  as a full stochastic transition matrix achieved by assigning teleportation probabilities to each node (uniform by using a personalization vector  $p(u) = 1/|N|$ ). This fact helps bootstrapping the introduction of new nodes by using, for example, profile similarity-based trust prediction [9] to address the newcomer problem.

HITS operates on a small portion of the graph, which is typically query dependent. Query dependent subgraphs permit online computation of hub and authority scores. Our previous work [8] addressed online computation of HITS using context information in collaborative environments. Furthermore, hub and authority scores are available for each node. This not only helps in finding authoritative nodes, but also *coordinators* (i.e., hubs) distributing work in online help and support communities.

### 3 Dynamic Skill and Activity-Aware PageRank

Based on the previous discussions on techniques for expertise mining, we propose *DSARank* to account for ranking of experts in context networks. Such networks carry context information through tagged interaction links expressing the importance of a link for a specific collaboration context.

#### 3.1 Definition

The main idea of our approach is to use *preferences* based on a set of interaction metrics to calculate personalized PageRank vectors. We define *DSARank* as:

$$DSARank(u) = \overbrace{(1 - \alpha) \cdot \sum_{w_m \in W_{M'}} w_m p_m(u)}^{\text{context-sensitive personalization}} + \alpha \cdot \overbrace{\sum_{(v,u) \in E} W(v,u) DSARank(v)}^{\text{importance propagation}} \quad (3)$$

The approach as shown in Eq. 3 is straightforward. For simplicity assume – without considering more advanced interaction metrics – that preferences should be assigned for finding experts in context  $C1$ :

$$p(u) \equiv \begin{cases} 1 & , \text{ if } u \in g(C1) \\ 0 & , \text{ otherwise} \end{cases} \quad (4)$$

**Table 3.** Metrics and weights in interaction networks

Symbol	Description
$m$	A metric (measured value) in interaction networks. The set $M' = \{m_1, m_2, \dots, m_n\}$ defines interaction metrics with $M' \subseteq M$ .
$w_m$	Denotes the weight of a metric to assign preferences for specific metrics. If not explicitly specified, equally important metrics are assumed with $W_{M'} = \{w w_m = 1/ M' , \forall m \in M'\}$ . The sum of metric weights must be equal to 1.
$W(v, u)$	The weight of the edge $(v, u) \in E$ calculated as $W(v, u) = 1/\text{outdegree}(v)$ .

$$I(u) \triangleq p_1(\text{subgraph } C1) + PR(\text{network}) + p_2(\text{subgraph } C2) + PR(\text{network})$$

**Fig. 2.** Expertise ranking using DSARank:  $I(u)$  is the importance of  $u$  given context networks  $C1$  and  $C2$ . Context-sensitive ranking scores are composed as the linear sum of personalized PageRank vectors.

Preferences are thereby given to those users who have interacted in a particular context  $C1$  with other users depicted as the subgraph  $g(C1)$ . For example, users have performed tasks in  $C1$  or have been involved in discussions related to  $C1$ . The problem with this naïve approach becomes evident when computing Eq. 3 for different preferences – say  $g(C1)$  and  $g(C2)$  or some other combinations of context networks  $C1 \dots Cn$ . The demanded set of skills and expertise of users in expert finder applications is typically only known at runtime (the *query* specified by the expert seeker). The online computation of Eq. 3 is, however, not feasible for large-scale interaction networks requiring up to several hours to calculate the importance of users.

We propose the PageRank linearity theorem to solve the problem of context-sensitive expertise mining. The linearity theorem [10,11] is defined as:

**Theorem 1 (Linearity).** *For any personalization vectors  $p_1, p_2$  and weights  $w_1, w_2$  with  $w_1 + w_2 = 1$ , the following equality holds:*

$$PPV(w_1p_1 + w_2p_2) = w_1PPV(p_1) + w_2PPV(p_2) \tag{5}$$

The above equality states that personalized PageRank vectors  $PPV$  can be composed as the weighted sum of PageRank vectors. Fig. 2 illustrates this concept where  $I(u)$  denotes the importance of  $u$  in the context networks  $C1$  and  $C2$ . The linearity theorem is a very important concept for creating context-sensitive expertise scores because it enables the *online* combination of PageRank vectors based on the expert seeker’s query preferences. Based on Eq. 5, Eq. 3 can be rewritten as:

$$DSARank^Q = \sum_{C \in Q} w_C \overbrace{DSARank(w_{m1}p_{m1} + w_{m2}p_{m2} \dots + w_{mn}p_{mn})}^{\text{context-sensitive importance}} \tag{6}$$

with  $\sum_m w_m = 1$  (see also Table 3) and  $\forall m : \sum_m p_m = 1$ .

### 3.2 Ranking Algorithm

Algorithm 1 shows the steps to compute context-sensitive importance scores as defined in Eq. 6. The steps 2-7 are used to assign metric values to the personalization vector  $p(u)$  by invoking the function `UpdatePrefVector`. The detailed

definition of interaction-based metrics follows in the next section. Steps 8-16 show the iterative computation of importance scores. The temp vector  $s(u)$  (see line 11) holds scores that are updated to  $DSARank$  scores in line 14 using the PageRank teleportation parameter  $\alpha$ .

```

input : Graph  $G$  and metrics  $M'$ 
output: Expertise scores  $DSARank$ 
1 /* initialize  $DSARank_1$  */
2 for Node  $u$  in  $G$  do
3   for Metric  $m$  in  $M'$  do
4     | UpdatePrefVector( $p(u), m$ )
5   end
6    $DSARank_1(u) \leftarrow p(u)$ 
7 end
8 while not converged do
9   for Node  $u$  in  $G$  do
10    | for Node  $v$  with  $(v, u) \in E$  do
11      |  $s(u) \leftarrow DSARank_t(u) + W(v, u)DSARank_t(v)$ 
12    | end
13    | /* update expertise scores */
14    |  $DSARank_{t+1}(u) \leftarrow (1 - \alpha) \cdot p(u) + \alpha \cdot s(u)$ 
15  | end
16  | /* Normalize rankings and test for convergence */
17 end
18 return  $DSARank$ 

```

**Algorithm 1.** Calculating context-sensitive importance scores

Convergence is reached once the change in ranking scores does not exceed a predefined threshold  $\epsilon$ :

$$\|DSARank_{t+1} - DSARank_t\| \leq \epsilon \quad (7)$$

While Algorithm 1 shows the offline mining procedure, the actual context-dependent ranking scores ( $DSARank^Q$ ) are aggregated based on the expert seeker's query preferences.

Algorithm 2 shows the composition of expertise ranking scores based on a set of context parameters (e.g., keywords, context identifiers, or context topic descriptions) formulated as an expert query  $Q$ . The weight  $w_C$  (line 3) is the preference for a specific context parameter. For example, the preference of the expert seeker that a person (expert) has stronger expertise in a specific topic. Again, the weights must satisfy  $\sum_C w_C = 1$ . This is due to the fact that the personalization vector  $\mathbf{p}$  has to represent a valid probability distribution.

```

input : Query  $Q$  specifying context parameters
output: Composite expertise scores
1 for Node  $u \in G$  do
2   for Context parameter  $C$  in  $Q$  do
3      $DSARank^Q(u) \leftarrow DSARank^Q(u) + w_C DSARank^C(u)$ 
4   end
5 end
6 return  $DSARank^Q$ 

```

**Algorithm 2.** Calculating composite expertise scores

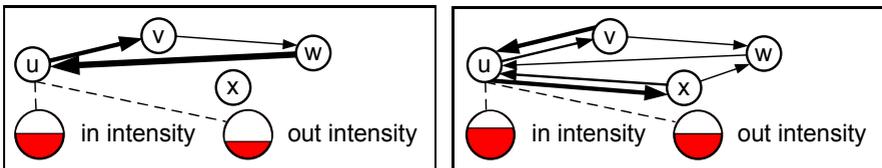
Algorithm 1 is a generic approach addressing the requirements of context-sensitive expertise mining. The key point is to define a set of metrics to capture dynamic user interests and context-dependent expertise. We have defined a broad range of context-sensitive link metrics in [12]. In the following section, we focus on a subset of metrics and show their impact in the evaluation section.

## 4 Interaction Metrics

The first metric defines the context (topic) sensitive interest (or focus) of a user. This can be measured as the *intensity* of a user interacting in a specific context compared to the user's focus on other contexts. Fig. 3 illustrates the basic idea of this approach. The interest and skills of people typically evolve over time. A collaboration partner may change his/her interaction behavior by focusing on, for example, a new topic of interest. The presented metric captures interaction intensities of all nodes in a given context. As a first step we create context dependent subgraphs as illustrated by Fig. 3. We assume that users tag their interactions (e.g., exchanged emails, delegated tasks, etc.) using a predefined set  $T = \{t_1, t_2, \dots, t_n\}$  of tags. Each tag appears with a certain frequency  $f(t_n)$ . Based on a set of accumulated tags between  $v$  and  $u$  we calculate a weighted edge  $(v, u) \in E$ :

$$W_C(v, u) \propto \frac{f(t_C) + \gamma}{\sum_{i=1..n} (f(t_{C_i}) + \gamma)} \quad (8)$$

The edge weight is essentially based on the number of tags associated by  $v$  when interacting with  $u$ . We apply a simple yet effective smoothing approach



**Fig. 3.** Weighted context network measuring interaction intensity of nodes. Nodes interact in different contexts with different intensities.

called *Lidstone smoothing* with  $0 < \gamma < 1$ , which is a well known technique in information retrieval to account for missing or misplaced tags. Using Eq. 8, various exchanged (tagged) resources influence the interaction weight between users. We now turn to the definition of a metric to measure the *intensity* of interactions of a specific user. For example (see Fig. 3),  $u$ 's incoming/outgoing links within certain subgraphs  $g$  are associated with different weights. Let us define the set of all links  $L(u) = \text{outlinks}(u) \cup \text{inlinks}(u)$  originating from  $u$  ( $\text{outlinks}(u)$ ) and pointing to  $u$  ( $\text{inlinks}(u)$ ).

**Definition 1 (Interaction Intensity).** *Let us define the interaction intensity of  $u$  over a link  $l$  as:*

$$i_C(l; u) = \left[ \frac{w(l; C)}{\sum_{l' \in L(u)} w(l'; C)} \right] \times |l| \tag{9}$$

The weight  $w(l; C)$  is based on the smoothed tag-based context weight using Eq. 8. The term in square brackets in Eq. 9 is the relative weight of the link  $l$  among all links attached to  $u$ . This is multiplied times  $|l|$ , the number of interactions associated with  $l$  (e.g., count of delegated tasks). Put in simple words, Eq. 9 combines three factors: *relevance of link for a given context*  $w(l; C)$ , *importance of link* ‘[...]’  $\times$  *strength of link*  $|l|$ .

Next, we define two further metrics based on the definition of  $i_C(l; u)$ .

**Definition 2 (Interaction Intensity Level).** *We define the interaction intensity level  $IIL_C$  of  $u$  as:*

$$IIL_C(u) = \left[ \beta^2 \left( \sum_{l \in \text{outlinks}(u)} i_C^{out}(l; u) \right)^2 + (2 - \beta)^2 \left( \sum_{l \in \text{inlinks}(u)} i_C^{in}(l; u) \right)^2 \right]^{(1/2)} \tag{10}$$

Eq. 10 calculates the overall interaction intensity of  $u$  in a given context  $C$ .  $IIL_C$  basically depicts the user’s interest and *activity level* in a specific context  $C$ . The factor  $\beta \in [0, 2]$  allows  $IIL_C$  to be biased towards  $i_C^{out}(l; u)$  or  $i_C^{in}(l; u)$ , where 1 means no bias, i.e., equal importance for in-/out intensities. Biasing  $IIL_C$  is only done for *all* nodes in the network to have valid probabilities for  $\mathbf{p}$ .

**Definition 3 (IIL Imbalance).** *Let us define  $imb_C(IIL) \in [-1, 1]$  as:*

$$imb_C(IIL) = \frac{\sum_{l \in \text{inlinks}(u)} i_C^{in}(l; u) - \sum_{l \in \text{outlinks}(u)} i_C^{out}(l; u)}{\sum_{l \in L(u)} i_C(l; u)} \tag{11}$$

$IIL$  imbalance  $imb(IIL)$  is a useful measure to determine two extreme values: *Passive* involvement of users in interactions such as ‘observers’ or ‘monitors’ resulting in  $imb(IIL)$  being equal to 1. *Active* involvement of users with the extreme case of all user interactions being outgoing (none of the interactions is reciprocated by others), thus  $imb(IIL) = -1$ . This metric is mainly used to

constraint and filter relevant users, for example, to rank only those users within a range  $0.1 < imb(IIL) < 0.9$ .

We defined the previous (intensity) metrics to model a user’s interest to interact with its neighbors in a particular context. These intensity metrics do not account for importance propagation within a contextual (sub)graph  $g(C_n)$ . For this purpose we define a metric to model interaction behavior and importance propagation in context networks. A user  $v$  may regard some interactions with  $u$  as important in a specific context. As mentioned before, the relevance of a link in a context  $C_n$  is proportional to the frequency of applied tags (cf. Eq. 8).

**Definition 4 (Skill and Expertise Fingerprints).** *Let us define the expertise of  $u$  in contextual subgraphs as:*

$$SE_C(u) = \sum_{(v,u) \in E} \left( \frac{W_C(v,u)}{\sum_{(v,w) \in E} W_C(v,w)} \right) SE_C(v) \quad (12)$$

We term the propagation of importance in contextual subgraphs (i.e., context networks) as Skill and Expertise  $SE_C(u)$  fingerprints.  $SE_C(u)$  can be interpreted as the *expected informedness* of a user  $u$  in context  $C$  to denote the chance of  $u$  receiving an item of information or that  $v$  will contact  $u$  in a specific context  $C$ .

## 5 Evaluation

To evaluate our ranking approach, we use a real email dataset because at the time of writing this paper, we did not have access to data generated by HPS-based crowdsourcing applications. To test the effectiveness of *DSARank*, we establish a set of ranking metrics to compare *DSARank* with the standard PageRank (as defined by Eq. 1) algorithm to obtain baseline results. In addition we select concrete examples of ranking results, for example a concrete collaboration context, and analyze how well these results reflect the actual expertise (or knowledge) of people. For evaluation, we use a set of well established ranking metrics.

### 5.1 Evaluation Metrics and Ranking Parameters

**Kendall’s  $\tau$ .** We use the definition of Kendall’s  $\tau$  by [13]. Consider the pairs of vertices  $v, w$ . The pair is *concordant* if two rankings agree on the order, *disconcordant* if both rankings disagree on the order, *exact-tie* if the exact ranking does not order the pair, and *approximate-tie* if the approximate ranking does not order the pair. Denote the number of these pairs by  $c, d, e$ , and  $a$ , respectively. Given the total number of pairs as  $m_p = \frac{n(n-1)}{2}$ ,  $n = |N|$ , then Kendall’s  $\tau \in [-1, 1]$  is defined as:  $\tau = (c - d) / \sqrt{(m_p - e)(m_p - a)}$ . Kendall’s  $\tau$  helps to understand whether two algorithms are rank-similar. In other words, if  $\tau$  equals 1, there are no cases where the pair  $v, w$  is ranked in a different order.

**Top- $k$  Set Metrics.** The overlap similarity  $OSim(T_{k1}, T_{k2})$  of the top- $k$  sets  $T_{k1}$  and  $T_{k2}$  ranked by *DSARank* and PageRank is defined as  $OSim(T_{k1}, T_{k2}) = \frac{|T_{k1} \cap T_{k2}|}{k}$  [10].

**Ranking Parameters.** We set the *IIL* parameter  $\beta$  to 1.2; thereby assigning a bias to out-intensities. We use a filter of  $-0.9 < imb(IIL) < 0.9$ . If  $imb(IIL)$  of a user is not within this range, we ‘downgrade’ the user’s  $IIL_C(u)$  to 0 because links between users might be irrelevant if *IIL* is strongly imbalanced.

**Teleportation Parameter  $\alpha$ .** There is some controversy whether the same parameter value as applied in the Web graph is suitable for different kind of networks (e.g., see [14] for a discussion on the damping factor in citation networks). In this work, we experimented with different values for  $\alpha$ . Low  $\alpha$  (e.g.,  $\alpha < 0.45$ ) results in unstable ranking results so that the proposed algorithm delivers quite different results compared to the baseline algorithm. In all experiments we use a factor of  $\alpha = 0.85$  yielding good results as discussed in the following.

## 5.2 Dataset Properties

We use the Enron email interaction network<sup>2</sup>, which is to our knowledge the only substantial collection of real email messages publicly available. The entire email graph comprises about 255635 messages and nearly 1647224 users (i.e., recipients of messages).

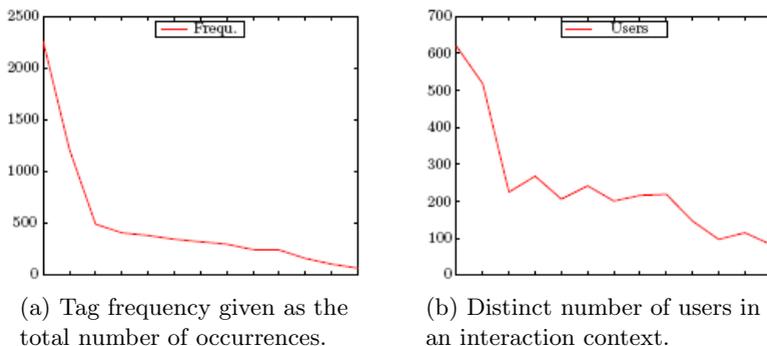
**Tagged Message Corpus.** A subset of messages of the entire message corpus was labeled by UC Berkeley’s Natural Language Processing group. These tags were applied to about 1700 messages. The tagged message corpus serves well to test our context-based ranking approach. Different categories of tags were applied to interaction links comprising messages between people with the focus on business-related emails. 13 categories or *contexts* are available (Table 4). Tag statistics are depicted by Fig. 4.

**Table 4.** Primary categories in labeled interaction graph. The index establishes the correspondence to the tag-statistics in Fig. 4 (horizontal axis).

ID	Index	Description	ID	Index	Description
3.1	2	Regulations and regulators (includes price caps)	3.8	9	Internal company operations
3.2	5	Internal projects – progress and strategy	3.9	11	Alliances / partnerships
3.3	3	Company image – current	3.10	4	Legal advice
3.4	6	Company image – changing / influencing	3.11	13	Talking points
3.5	10	Political influence / contributions / contacts	3.12	8	Meeting minutes
3.6	1	California energy crisis / California politics	3.13	12	Trip reports
3.7	7	Internal company policy			

<sup>2</sup> Email Dataset: [http://bailando.sims.berkeley.edu/enron\\_email.html](http://bailando.sims.berkeley.edu/enron_email.html)

**Applied Expansion and Filtering.** We expand the subset of labeled messages by searching the entire email-message corpus for related messages. For example, given a labeled message, we search for messages which are most likely related to the labeled message; for example, in *reply-to* or *forwarded* messages. Thereby, we expand the existing labeled message corpus by adding 5248 related messages. However, some messages are simply ‘broadcast’ messages (e.g., announcements or periodic auto-generated messages from a person), which we filter out because these messages might distort ranking results. In addition, sender and recipient of messages must be real people (e.g., we filter messages *from* and *to* distribution lists) because we focus on importance rankings of people.

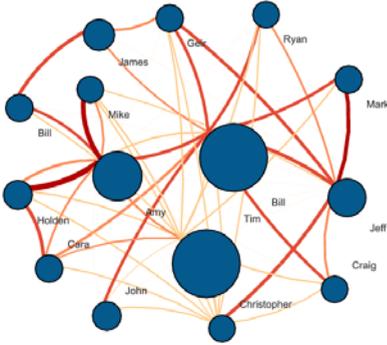


**Fig. 4.** Tag statistics in email interactions

### 5.3 Basic Results

As a first example, we select the subgraph for ID 3.6 (i.e., the specific interaction context) and rank all users. The results are detailed in Fig. 5 by a graph visualization and a detailed description in the table. The goal of testing *DSARank* in context networks is to verify whether the obtained ranking results contain the most knowledgeable people in a specific context. Using *DSARank* in category 3.6, we verified that the top-ranked users were indeed key people in the given context. Not only the number of inbound interactions with important users plays a key role, but also *interaction intensity* in a given context dependent subgraph. It is well possible that users get in some cases substantially promoted (i.e., higher rank) because they interact with important users in a given context with high intensity. *DSARank* provides accurate results since users are not ranked in a single context. The primary goal of this work was not to understand in detail the roles and whereabouts of the email dataset. Therefore, we do not provide the details in this paper. For validation of our results, we provide details<sup>3</sup> about the user identities as well as descriptions of specialism (where available).

<sup>3</sup> Email Mining: <http://www.infosys.tuwien.ac.at/staff/dschall/email/>



ID	scoreA	rankA	scoreB	rankB	<i>IIL</i>	<i>imb</i>
37	0.109475	1	0.004121	21	7.31	-0.81
8	0.102375	2	0.020758	1	5.13	0.11
90	0.043071	3	0.008326	9	1.10	0.08
253	0.029936	4	0.001733	170	2.07	-0.85
347	0.020443	5	0.001665	282	1.39	-0.87
92	0.016677	6	0.003912	23	0.39	0.82
152	0.016375	7	0.013148	2	1.16	1.00
47	0.014248	8	0.003593	27	0.66	0.41
29	0.014195	9	0.005415	16	1.14	1.00
14	0.014084	10	0.010911	4	2.27	1.00

<b>User ID</b> (ordered by rank)	37	8	90	253	347	92	152	47	29	14
<b>User Name</b> (see graph)	Tim	Bill	Amy	Jeff	James	Holden	John	Cara	Bill	Mark

**Fig. 5.** Filtered interaction graph of selected collaboration context and results: columns *scoreA/rankA* show results by *DSARank* and *scoreB/rankB* by unbiased PageRank

### 5.4 Comparison of Composite Contexts

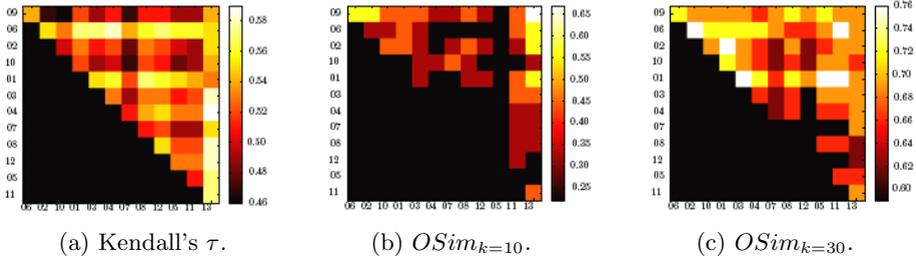
Here we analyze the impact of composite context-sensitive ranking results. We rank in different subgraphs and combine results using Eq. 6 to create composite *DSARank* scores. Each context-dependent result vector  $DSARank^{C1}$  and  $DSARank^{C2}$  is combined with  $w_{C1} = w_{C2}$ . To create PageRank scores, we use the entire interaction graph  $G$  to create the vector  $PR$ . Furthermore, let us define a metric to compare the ‘dominance’ of context-dependent subgraphs given the set of contexts  $C_N = \{C1, C2, \dots, Cn\}$ .

We calculate the intensity of all users  $U(C)$  in  $g(C)$  as subgraph intensity:

$$i(g(C)) = \frac{1}{|U(C)||C_N|} \sum_{i=1..n} \sum_{u \in U(Ci)} IIL_{Ci}(u) \tag{13}$$

Each subfigure in Fig. 6(a) shows the comparison of *DSARank* and PageRank ranking results using previously introduced ranking metrics. We start with the first Fig. 6(a) depicting Kendall’s  $\tau$ .

**Kendall’s  $\tau$ .** Fig. 6(a) visualizes the results to indicate whether there is a strong disagreement (low value for Kendall’s  $\tau$ ) when combining different contexts. On the other hand, if Kendall’s  $\tau$  is equal to 1, then ranking considering context information does not change the ranking order (i.e., using context for expertise mining does not bring any additional benefit). Contextual graphs that share links with other context graphs (i.e., links with a shared set of tags) yield stronger agreements between *DSARank* and PageRank. This is the case in, for example, categories 3.6 and 3.1 (see IDs in Table 4), which share a high degree of links (tags). Thus, 3.6 and 3.1 become more dominant when combined with other contexts. It is therefore less likely that the order of rankings change.



**Fig. 6.** Comparison of composite contexts using Kendall's  $\tau$  and  $OSim$  metrics

On the other hand, if a context, for example category 3.13 (the corresponding subgraph) has few shared links with other contexts; and also low subgraph intensity (3.13 has the lowest subgraph intensity), then we observe also stronger agreements in rankings. This can be explained as the limited impact of low intensity contexts on changing the position of users within ranking results.

**Overlap Similarities  $OSim$ .** We compared  $DSARank$  and PageRank in terms of overlap similarities. In Figure 6(b) and 6(c) we show the visualizations of the results. By comparing the top-10 segment of ranked users (Figure 6(b)), we see higher overlap similarities between high-intensity contexts, for example, the context pairs (3.9, 3.6), (3.9, 3.2). Low intensity contexts such as 3.13 combined with, for example (3.13, 3.9), yields also high similarities. The top-30 segment (Figure 6(c)) shows stronger similarities in 3.1 as well as 3.6; both contexts have many shared links with other contexts.

To conclude our experiments, PageRank provides a valuable method for expertise ranking. This has also been confirmed by other research (e.g., see [5]). However, ‘context’ is still a key issue in expertise ranking as the level of user expertise varies depending on context. Based on discussions of Kendall's  $\tau$  and  $OSim$  to compare ranking approaches, parameterization (i.e., personalization) through interaction metrics is important to model user interest, activity level, and expertise in context networks.

## 6 Related Work

Task-based platforms on the Web allow users to share their expertise [15]. Also users offer their expertise by helping other users in forums or Q/A communities [6,16]. By analyzing email conversations [17], the authors studied graph-based algorithms such as HITS and PageRank to estimate the expertise of users. The authors in [18] followed a graph-entropy model to measure the importance of users. In [19], an email analysis in enterprises, defining information flow metrics in the social interaction graph was presented. The work by [5] followed a graph-based approach and applied HITS as well as PageRank in online communities (i.e., a Java question and answer forum). The authors in [20] studied minimum spanning tree algorithms in social networks.

While the above cited works attempted to model the importance of users based on interactions; they did not consider that interactions typically take place in different *contexts*. Approaches for calculating personalized PageRank scores [10,11] were introduced to enable topic-sensitive search on the Web. The novelty of our approach is that expertise mining is performed by considering contextual link information.

## 7 Conclusion

In this work we have motivated the need for context-sensitive expertise mining in open collaboration environments. As a first step, we compared popular and well established methods including PageRank and HITS. Most existing approaches utilizing these algorithms, however, do not consider context information for expertise mining. Here we systematically derived a context-sensitive expertise mining technique and algorithm called *DSARank*. Furthermore, we defined a set of metrics to measure collaboration characteristics to capture user-interest and contextual skills.

We are currently working on HPS testbeds to test different interaction models in crowdsourcing applications. Interactions in the testbed are based on Web service technology such as SOAP-based messages exchanged between HPSs and dynamically routed by means of delegations in HPS networks. This will help to perform more experiments in realistic service-oriented crowdsourcing scenarios.

## References

1. Schall, D., Truong, H.L., Dustdar, S.: Unifying Human and Software Services in Web-Scale Collaborations. *IEEE Internet Computing* 12(3), 62–68 (2008)
2. Becerra-Fernandez, I.: Searching for experts on the Web: A review of contemporary expertise locator systems. *ACM Trans. Inter. Tech.* 6(4), 333–355 (2006)
3. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report (1998)
4. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. *J. ACM* 46(5), 604–632 (1999)
5. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: *WWW*, pp. 221–230. ACM, New York (2007)
6. Jurczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: *CIKM 2007*, pp. 919–922. ACM, New York (2007)
7. Berkhin, P.: A survey on pagerank computing. *Internet Mathematics* 2, 73–120 (2005)
8. Schall, D., Skopik, F., Dustdar, S.: Trust-based discovery and interactions in mixed service-oriented systems. Technical Report TUV-1841-2010-01 (2010)
9. Skopik, F., Schall, D., Dustdar, S.: Start trusting strangers? bootstrapping and prediction of trust. In: Vossen, G., Long, D.D.E., Yu, J.X. (eds.) *WISE 2009*. LNCS, vol. 5802, pp. 275–289. Springer, Heidelberg (2009)
10. Haveliwala, T.H.: Topic-sensitive pagerank. In: *WWW*, pp. 517–526 (2002)
11. Jeh, G., Widom, J.: Scaling personalized web search. In: *WWW*, pp. 271–279. ACM, New York (2003)

12. Schall, D.: Human Interactions in Mixed Systems - Architecture, Protocols, and Algorithms. PhD thesis, Vienna University of Technology (2009)
13. Fogaras, D., Csalogany, K., Racz, B., Sarlos, T.: Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Mathematics* 2(3), 333–358 (2005)
14. Maslov, S., Redner, S.: Promise and pitfalls of extending google’s pagerank algorithm to citation networks. *J. Neurosci.* 28(44), 11103–11105 (2008)
15. Yang, J., Adamic, L., Ackerman, M.: Competing to share expertise: the taskcn knowledge sharing community. In: *Conference on Weblogs and Social Media* (2008)
16. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: *WSDM 2008*, pp. 183–194. ACM, New York (2008)
17. Dom, B., Eiron, I., Cozzi, A., Zhang, Y.: Graph-based ranking algorithms for e-mail expertise analysis. In: *DMKD 2003*, pp. 42–48. ACM, New York (2003)
18. Shetty, J., Adibi, J.: Discovering important nodes through graph entropy the case of enron email database. In: *LinkKDD 2005*, pp. 74–81. ACM, New York (2005)
19. Karagiannis, T., Vojnovic, M.: Email Information Flow in Large-Scale Enterprises. Technical report, Microsoft Research (2008)
20. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: *KDD 2009*, pp. 467–476. ACM, New York (2009)