ISWC 2008
The 7th International Semantic Web Conference

Malgorzata Mochol
Anna V. Zhdanova
Lyndon J. B. Nixon
John Breslin
Axel Polleres

Personal Identification and Collaborations: Knowledge Mediation and Extraction (PICKME 2008)

October 27, 2008

The 7th International Semantic Web Conference
October 26 – 30, 2008
Congress Center, Karlsruhe, Germany

ISWC 2008

Platinum Sponsors

Ontoprise

Gold Sponsors

BBN
eyeworkers
Microsoft
NeOn
SAP Research
Vulcan

Silver Sponsors

ACTIVE
ADUNA
Saltlux
SUPER
X-Media
Yahoo

The 7th International Semantic Web Conference
October 26 – 30, 2008
Congress Center, Karlsruhe, Germany

ISWC 2008

Organizing Committee

General Chair
Tim Finin (University of Maryland, Baltimore County)

Local Chair
Rudi Studer (Universität Karlsruhe (TH), FZI Forschungszentrum Informatik)

Local Organizing Committee
Anne Eberhardt (Universität Karlsruhe)
Holger Lewen (Universität Karlsruhe)
York Sure (SAP Research Karlsruhe)

Program Chairs
Amit Sheth (Wright State University)
Steffen Staab (Universität Koblenz Landau)

Semantic Web in Use Chairs
Mike Dean (BBN)
Massimo Paolucci (DoCoMo Euro-labs)

Semantic Web Challenge Chairs
Jim Hendler (RPI, USA)
Peter Mika (Yahoo, ES)

Workshop chairs
Melliyal Annamalai (Oracle, USA)
Daniel Olmedilla (Leibniz Universität Hannover, DE)

Tutorial Chairs
Lalana Kagal (MIT)
David Martin (SRI)

Poster and Demos Chairs
Chris Bizer (Freie Universität Berlin)
Anupam Joshi (UMBC)

Doctoral Consortium Chairs
Diana Maynard (Sheffield)

Sponsor Chairs
John Domingue (The Open University)
Benjamin Grosof (Vulcan Inc.)

Metadata Chairs
Richard Cyganiak (DERI/Freie Universität Berlin)
Knud Möller (DERI)

Publicity Chair
Li Ding (RPI)

Proceedings Chair
Krishnaprasad Thirunarayan (Wright State University)

Fellowship Chair
Joel Sachs (UMBC)

Table of contents

Preface

The Semantic Web, Social Networks and other emerging technology streams promise to enable finding experts more efficiently on a Web scale across boundaries. To leverage synergies among these streams, the ***ExpertFinder Initiative***[1] started in 2006 with the aim of devising vocabularies, rule extensions (for e.g. FOAF and SIOC) and best practices to annotate and extract expertise-relevant information from personal and organizational web pages, blogs, wikis, conferences, publication indexes, etc. Following two previous workshops - ***EFW***[2] and ***FEWS***[3] - we solicit new research contributions from the Semantic Web community towards the tasks of formally representing and reusing knowledge of skills and collaborations on the Web and consequently finding people according to their expertise.
The goal of ***PICKME2008*** is to discuss:

- the feasibility of a Web-scale infrastructure for the creation, publication and use of semantic descriptions of experts and their collaborations on the Web,
- concrete application scenarios such as group management, disaster response, recruitment, team building, problem solving and on-the-fly consultation4,
- enabling technologies such as annotation, knowledge extraction, ontology engineering, reasoning, ontology mediation, social network and interaction analysis.

PICKME2008 welcomes all research contributions that address one or more of the following topics:

- Specification of vocabularies and reuse of existing standards/taxonomies to describe experts to capture knowledge about people, their expertise and collaborations with other people,
- Extraction of descriptions of persons and collaborations from loosely structured data (e.g. Web pages) and databases,
- Use of microformat to express and extract knowledge about persons and collaborations,
- International and cross-organizational heterogeneity issues in personal descriptions,
- Algorithms for expert & expertise finding and recommendation (e.g. mining of social networks),
- Expressivity extensions (in logics, rules) to support expertise extraction from knowledge about collaborations,
- Tools for the intuitive creation and maintenance of personal and organizational descriptions and associated rules,
- Web infrastructures for the publication and sharing of personal and organizational descriptions (storage, access, querying, rule execution, coordination, communication),
- Extension of collaborative tools, e.g. blogs and wikis, to capture knowledge about persons and collaborations,
- Security, trust and privacy aspects of expert & expertise finding, and
- Deployment of these areas in business scenarios and requirements for the industrial uptake of these applications.

The workshop organising committee thanks the ISWC2008 organisers who helped us tremendously by caring about most of the logistics and overall technical issues. We are grateful to our keynote speaker Dr. Ora Lassila and to the members of the program committee who completed the paper reviews in a short turnaround time.

Enjoy PICKME2008!

Your Organising Committee

[1] http://expertfinder.info/
[2] http://expertfinder.info/efw2007
[3] http://expertfinder.info/fews2007
[4] Such use cases can be found at: http://wiki.foaf-project.org/ExpertFinderUseCases

Program Committee

- Witold Abramowicz, Poznan University of Economics, Poland;
- Diego Berrueta, CTIC Foundation, Spain;
- Chris Bizer, Freie Unversität Berlin;
- Aleman-Meza Boanerges, LSDIS Lab, University of Georgia, USA;
- Harold Boley, NRC Institute of Information Technology, Canada;
- Irene Celino, CERIEL Politecnico di Milano, Italy;
- Asunción Gómez-Pérez, Universidad Politecnica de Madrid (UPM), Spain;
- Tom Heath, Talis Information Lt, UK;
- Alain Leger, France Telekom, France;
- Ning Li, University of Surrey, UK;
- Benjamin Nowack, semsol web semantics, Germany;
- Charles Petrie, Stanford Unviersity, USA;
- Robert Tolksdorf, Freie Universität Berlin, Germany;
- Holger Wache, University of Applied Sciences Northwestern Switzerland (FHNW), School of Business, Switzerland.

Organising Committee

- John Breslin, DERI, NUI Galway, Ireland
- Malgorzata Mochol, Free University of Berlin, Germany
- Lyndon J. B. Nixon, Free University of Berlin, Germany
- Axel Polleres, DERI, NUI Galway, Ireland
- Anna V. Zhdanova, ftw. Forschungszentrum Telekommunikation Wien, Austria

Keynote

„Does Your Mobile Device Need to be Socially Aware? A Semantic Web Perspective“ - Ora Lassila

Abstract

Personal Information Management (PIM) has evolved from simple maintenance of user-created data to access and management of data relevant and related to the user and the user's social network. Users also typically have multiple systems/devices they use for maintaining (and communicating with) their social network. In this talk I will discuss the need for pervasive awareness of the user's social context. Semantic Web technologies offer possibilities for a rich representation of social networks, in turn enabling new types of applications and functionality.

Biography

Ora Lassila is a Research Fellow at the Nokia Research Center in Cambridge (Massachusetts, USA). His research work focuses on the applications of Semantic Web technology to mobile devices and personal information management. Lassila pioneered the Semantic Web vision in the late 1990s, and has worked on many of the fundamental aspects of the technology. He holds a Ph.D. from Helsinki University of Technology."

Finding Experts By Semantic Matching of User Profiles

Rajesh Thiagarajan[1], Geetha Manjunath[2], and Markus Stumptner[1]

[1] Advanced Computing Research Centre, University of South Australia
{cisrkt,mst}@cs.unisa.edu.au
[2] Hewlett-Packard Laboratories, India
geetha.manjunath@hp.com

Abstract. Extracting interest profiles of users based on their personal documents is one of the key topics of IR research. However, when these extracted profiles are used in expert finding applications, only naive text-matching techniques are used to rank experts for a given requirement. In this paper, we address this gap and describe multiple techniques to match user profiles for better ranking of experts. We propose new metrics for computing semantic similarity of user profiles using spreading activation networks derived from ontologies. Our pilot evaluation shows that matching algorithms based on bipartite graphs over semantic user profiles provide the best results. We show that using these techniques, we can find an expert more accurately than other approaches, in particular within the top ranked results. In applications where a group of candidate users need to be short-listed (say, for a job interview), we get very good precision and recall as well.

1 Introduction

The problem of finding experts on a given set of topics is important for many lines of business e.g., consulting, recruitment, e-business. In these applications, one common way to model a user is with a *user profile* which is a set of topics with weights determining his level of interest. When used for personalization, these user profiles matched with a retrieved document (may be a search result) for checking its relevance to him. A similar matching technique can be used for expert finding as well - wherein we first formulate the requirement (query) as an expected profile of the expert who is sought after. Expert finding is then carried out by matching the query profile with the available/extracted expert user profiles.

In the above context, automatic extraction of topics of expertise (interest) of a person based on the documents authored (accessed) by the person through information extraction techniques is well known. However, when these extracted profiles are used for expert finding, the profile matching is often carried out by applying traditional content matching techniques which miss most potential candidates if the query is only an approximate description of the expert (as is usually the case). In this paper, we propose and evaluate multiple approaches for semantic matching of user profiles to enable better expert-finding in such cases.

Let us briefly look at the challenges in comparing user profiles. User profiles are generally represented in the *bag-of-words* (BOW) format - a set of weighted terms that describe the interest or the expertise of a user. The most commonly used content matching technique is cosine similarity - cosine between the BOW vector representing the user profile and that of the document to match. Although this simple matching technique suffices in a number of content matching applications, it is well known that considering just the words leads to problems due to lack of semantics in the representation. Problems due to polysemy (terms such as *apple, jaguar* having two different meanings) and synonymy (two words meaning almost the same thing such as *glad* and *happy*) can be solved if profiles are described using semantic concepts instead of words. Once again simple

matching techniques can be used on these *bags-of-concepts* (BOC). However, these approaches fail to determine the right matches when there is no direct overlap/intersection in the concepts. For example, do two users with *Yahoo* and *Google* in their respective profiles have nothing in common? There does seem to be an intersection in these users' interests for Web-based IT companies or web search tools! Such overlaps are missed as current approaches work under the assumption that the profile representations (BOW) contain all the information about the user. As a result, relationships that are not explicit in the representations are usually ignored. Furthermore, these mechanisms cannot handle user profiles that are at different levels of granularity or abstractions (e.g., *jazz* and *music*) as the implicit relationship between the concepts is ignored.

In this paper, we solve the above issues in user profile matching through effective use of ontologies. We define the notion of semantic similarity between two user profiles to consider inherent relationships between concepts/words appearing in their respective BOW representation. We use the process of *spreading* to include additional related terms to a user profile by referring to an ontology (Wordnet or Wikipedia) and experiment with multiple techniques to enable better profile matching. We propose simple metrics for computing similarity between two user profiles with ontology-based Spreading Activation Networks (SAN). We evaluate multiple mechanisms for extending user profiles (set and graph based spreading) and semantic matching (set intersection and bipartite graphs) of profiles. We show the effectiveness of our user profile matching techniques for accuracy in expert-ranking as well as candidate selection. From a given set of user profiles, our bipartite-graph based algorithms can accurately spot an expert just within its top three ranks. In applications where a group of candidate users need to be found (for a job interview), we get very good precision and recall as well.

The organization of the rest of this document is as follows. We describe different related research efforts for profile building and ontology-based semantic matching techniques in section 2 followed by a brief section giving some background and definitions needed to understand our solution. An overview of the our spreading process is presented in Section 4. We present our new similarity measures in Section 5. We describe our evaluation procedure for expert finding in Section 6 and share our improved results. We summarize our contributions and state possible future work in Section 7.

2 Related Work

Determining interest profiles of users based on their personal documents is an important research topic in information extraction and a number of techniques to achieve this have been proposed. Expert finding techniques that combine multiple sources of expertise evidence such as academic papers and social citation network have also bee proposed [1]. User profiles have been extracted using multiple types of corpora - utilizing knowledge about the expert in Wikipedia [2], analysing the expert's documents [3–5], and analysing openly accessible research contributions of the expert [6]. Use of Wikipedia corpus to generate semantic user profiles [7] have been seen. Pre-processing the profile terms by mapping terms to such ontology concepts prior to computing cosine similarity has been shown to yield better matching [3]. A number of traditional similarity measurement techniques such as the cosine similarity measure or term vector similarity [8, 9], Dice's coefficient [10] and Jaccard's index [11] are used in profile matching. For example, Jaccard's index is used in [2] to match expert profiles constructed using Wikipedia knowledge.This approach will not determine a semantic inexact match when there is no direct overlap in the concepts in the two user profiles. Use of knowledge obtained from

an ontology,in our solution, enables similarity checks when there are no direct overlaps between user profiles and, therefore, result in more accurate similarity measurements.

The problem of automated routing of conference papers to their reviewers is a somewhat related problem to that of expert finding. Most of the current approaches to that problem use a group of papers authored by reviewers to determine their user profile and perform routine content matching (similar to personalization) to determine whether a paper is fit to be reviewed by that user [12]. The expert finding task introduced by TREC 2005 [13] requires one to provide a ranked list of the candidate experts based on the web data provided. Our attempt is to handle the problem of choosing the best expert given a description of a hypothetical expert (set of topics with weights) and a set of user profiles of candidate experts.

Use of ontologies to derive new concepts that are likely to be of interest to the user through semantic spreading activation networks has been studied as well [14–17, 5]. Previous studies have shown that the spreading process improves accuracy and overcomes the challenges caused by inherent relationships and Polysemy in word sense disambiguation process [15, 16] and ontology mapping [17]. We use this spreading process to facilitate the semantic similarity computation. We build on the spreading process used in [5] to learn user preferences in order to drive a personalized multimedia search. The learning process utilizes ontologies as a means to comprehend user interests (in BOW format) and establishes the need to consider related concepts to improve search quality. While the results in [5] suggest that personalized search is of better quality in comparison to normal search, they do not show whether the consideration of related terms contributes to these improvements. On the other hand, we show that our spreading process indeed improves the accuracy of our new similarity measures and in the particular context of user profile matching.

A number of approaches have already been proposed to determine the similarity between two ontology concepts (or words). These determine similarity by: measuring the path distance between them [18], evaluating shared information between them [19], recursively matching sub-graphs [20], combining information from various sources [21], analysing structure of the ontology [22], and combining content analysis and web search [23]. A few other measures are evaluated in [24]. While all these approaches are only able to determine closeness between two concepts (or words), we compute similarity between two weighted sets of concepts (or words). One of our algorithms use the simple path measure described in [18] over a bipartite graph to determine such a set intersection.

We now compare with other works that use SAN based IR techniques. One of our similarity measures is similar to the one discussed in [25] but differs in the treatment of the results of the activation process. While the previous work utilizes the results of the activation to rank documents with respect to a query, our work maps an aggregate of the activation results to a similarity value. Knowledge from an ontology is used to extend the BOW with terms that share important relationships with original terms to improve document retrieval is presented in [4]. Our work on set spreading is somewhat similar to this but we further explore the notion of computing similarity by optimal concept matching in bipartite graphs and using SAN.

3 Background

In this section, we formally define and explain some terms used in the rest of the document.

Definition 1 (User Profile). *An user profile, u is a set of binary tuples $\{\langle t_1, w_1\rangle, \ldots, \langle t_n, w_n\rangle\}$ where t_i are the terms that describes the user and w_i denotes the importance of t_i in describing the user. We use $terms(u)$ to denote the set of terms t_i in the profile u.*

Cosine Similarity: The BOW representation is typically used for computing cosine similarity between the user profiles. If the vector representation of a user profile u_j is $\overrightarrow{V}(u_j)$ and the *Euclidean length* ($|\overrightarrow{V}(u_j)|$) of an entity u_j is $\sqrt{\sum_{i=1}^{n} w_i^2}$, the similarity of the entities u_j and u_k is

$$sim_{cos}(u_j, u_k) = cos(\overrightarrow{V}(u_j), \overrightarrow{V}(u_k)) = \frac{\overrightarrow{V}(u_j) \cdot \overrightarrow{V}(u_k)}{|\overrightarrow{V}(u_j)||\overrightarrow{V}(u_k)|} \quad (1)$$

Spreading: Spreading is the process of including the terms that are related to the original terms in an user profile by referring to an ontology. Let us study the earlier mentioned simple example of two users having *google* and *yahoo* in their profile in detail to understand the spreading process better.

Example 1. Consider computing the similarity of the following users
- $u_1 = \{\langle google, 1.0\rangle\}$, and
- $u_2 = \{\langle yahoo, 2.0\rangle\}$.

A simple intersection check between the profiles result in an empty set (i.e. $u_1 \cap u_2 = \emptyset$) indicating their un-relatedness (cosine similarity is 0). However, if we were to manually judge the similarity of these two users we would give it a value greater than 0. This is because we judge the similarity not just by considering the two terms from the profiles but also by considering the relationships that might exist between them due to our prior knowledge. We are able to establish the fact that both $google$ and $yahoo$ are search engine providers.

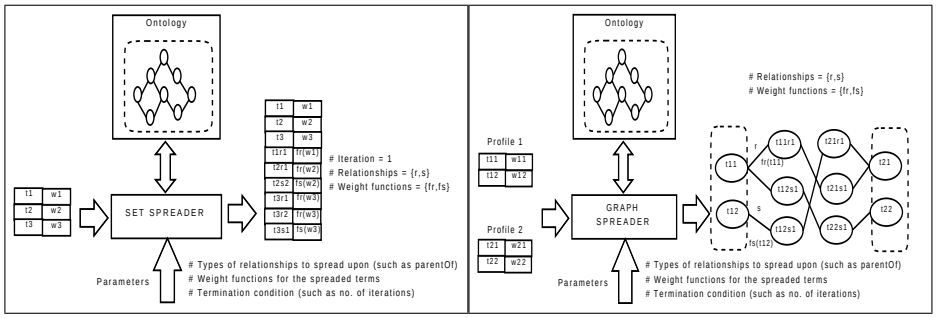
Now let us see the effectiveness of spreading in the similarity computation process in the same example. Spreading the profiles u_1 and u_2, by referring to Wikipedia parent category relationship, extends the profiles to
- $u_1' = \{\langle google, 1.0\rangle, \langle internet\ search\ engines, 0.5\rangle\}$, and
- $u_2' = \{\langle yahoo, 2.0\rangle, \langle internet\ search\ engines, 1.0\rangle\}$.

The simple intersection check results in a non-empty set (i.e. $u_1' \cap u_2' \neq \emptyset$) indicating their relatedness (cosine similarity is 0.2). The result of the spreading (i.e. the inclusion of the related term $internet\ search\ engines$) process makes sure that any relationship that exists between the profiles are taken into consideration.

4 Spreading to Create Extended User Profiles

In this section, we describe two techniques to compute and represent the extended user profiles (see example of section 3) using an ontology. An ontology $\mathcal{O}$ represents human knowledge about a certain domain as concepts, attributes and relationships between concepts in a well-defined hierarchy. It is usually represented as a graph where nodes are the concepts and edges are the relationship labelled with the type of relationship. For the purpose of profile spreading we assume that all the terms t_i describing an entity are mappable to concepts in a reference ontology. For example, all the terms t_i in a BOW representation of a user profile maps to a concept in the Wordnet ontology. Given a term t_i, the spreading process utilizes $\mathcal{O}$ to determine the terms that are related to t_i (denoted as $related_{\mathcal{O}}(t_i)$). Although spreading the profiles with related terms allows for

a comprehensive computation, uncontrolled addition of all the related terms leads to the dilution of the profiles with noise or unrelated terms. This dilution may have negative implications on the computation process where the similarity in the noise may contribute to the similarity values between entities. It is therefore desirable to have control over the types of relationships to be considered during this spreading process.



(a) Set Spreading (b) Graph Spreading

Fig. 1: Two Schemes for Profile Spreading

The weights of the new related terms are proportional to the weights of the original term as the weight w_i of a term t_i indicates the importance of the term within a user profile. However, during spreading the weights of the related terms should differ according to the semantics of the relationships on the edge. For example, spreading based on Wikipedia may be limited to only spreading along the parent categories. We therefore use a set of linear influence functions, one per relationship-type (role/property of an ontology), to control the spreading process. For example, a spreading process based on Wordnet limited to types synonym and antonym can have functions $t_{ij} = w_i \times 0.9$ and $t_{ij} = w_i \times -0.9$ respectively. We propose two schemes for representing the related terms post-spreading: extended set and semantic network.

4.1 Set Spreading

Depicted in Figure 1a, set spreading is a process of extending an user profile such that the related terms, which are determined with respect to an ontology, are just appended to the original set of terms. Set spreading is an iterative process. After each iteration, the related terms from the previous iterations are appended to the profile. The spreading process is terminated if there are no related terms to spread the profile with or after a fixed number of iterations.

4.2 Graph spreading

Shown in Figure 1b, graph spreading is the process where terms from two profiles and the related terms are build into a semantic network (SAN). Unlike set spreading, graph spreading preserves the relationship between a term in a profile and its related term in the form of a graph edge. This allows consideration of relationships based on their semantics on the same network. Graph spreading terminates like set spreading, or if there exists a path between every pair of the term nodes from the two profiles. This condition best suits the ontologies that have a top root element which subsumes the rest of the elements

in the ontology. For example, Wordnet based spreading can be tuned to employ this termination condition when path from individual terms to the root suffices to terminate the spreading. In less rigorous ontologies such as the Wikipedia category graph may not be able to support this condition as there may not be a single root. In such a case, the spreading process is terminated if there exists at least one path from every node that belongs to the smallest of the two profiles to the nodes in the other profile. We describe the complete details of the two spreading algorithms in our technical report [26].

5 Similarity Computation

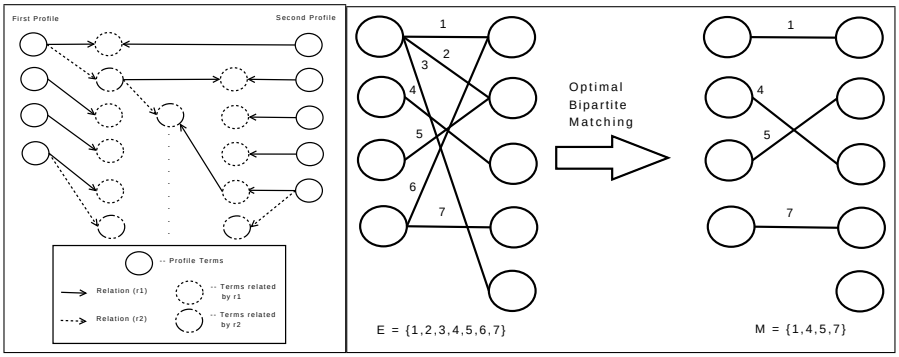
In this section, we describe the complete details of our variant metrics to compute semantic similarity using ontologies.

5.1 Set-based Measure

Set spreading process enriches the profiles by appending the related terms in order to capture all the relationships between the terms. For set spreading, the same cosine similarity technique defined in Equation 1 is applicable to compute similarity between the extended BOWs or BOCs. Set spreading-based similarity computation begins by measuring similarity of the original profiles, and proceeds by incrementally extending the profiles until termination while computing the similarity between profiles at every iteration.

5.2 SAN-based measure

This similarity computation metric is inspired by the abundant work that exists in the area of semantic search especially by techniques that process a SAN (e.g., [25, 15]). We focus on similarity computation techniques that use a SAN resulting from graph spreading process (see figure 2a for an overview of SAN structure). Following the construction of the semantic network the similarity values are computed either by reducing the graph to a bipartite graph or by activating the graph with an activation strategy. We have implemented both these techniques for evaluation. A brief introduction to the activation process is presented below. For a more detailed discussion the reader is pointed to [15].



(a) SAN Building Process (b) Bipartite Graph Matching (Hungarian Algorithm)

Fig. 2: SAN-based Similarity Computations

The SAN activation process is iterative. Let $A_j(p)$ denotes the activation value of node j at iteration p. All the original term nodes corresponding to the tuples in a user profile t_j take their term weights w_j as their initial activation value $A_j(0) = w_j$. The activation value of all the other nodes are initalized to 0. In each iteration,

- Every node propagates its activation to its neighbours.
- The propagated value is a function of the nodes current activation value and weight of the edge (see [15]) that connects them (denoted as $O_j(p)$).

After a certain number of iterations when the termination condition is reached, the highest activation value among the nodes that are associated with each of the original term node is retrieved into a set $ACT = \{act_1, act_2, ..., act_{n+m}\}$[3]. The aggregate of the these activation values can be mapped to the similarity between the profiles under the intuition that the nodes with higher activation values are typically the ones that have value contributions from both the profiles and hence should contribute more to similarity and vice versa. Therefore, the similarity value is the sum of the set ACT normalized to a value between 0 and 1. The SAN-based similarity between two profiles u_1 and u_2 where $max(ACT)$ is the highest activation value is

$$sim_{san}(u_1, u_2) = \frac{\sum_{\forall act_i \in ACT} act_i}{|ACT| \times max(ACT)} \quad (2)$$

5.3 Similarity Computation by Matching Bipartite Graph

A key insight here is that by omitting the intermediate related nodes and considering only the path length between the nodes representing the original profile terms, the semantic network can be converted to a bipartite graph (shown on the left side of Figure 2b). The nodes of the first profile and second profile are the two vertex sets of the bipartite graph where the edge denotes the length between the original term nodes as obtained from the semantic network. Once the bipartite graph is derived, we are able to apply standard algorithms for optimal matching of the bipartite graph. Our similarity measures based on optimal bipartite matching operates under the simple notion that the nodes with higher weights and that are closely located contribute more to the similarity of the entities and viceversa.

Each node v_i^u in the semantic network is a pair $\langle t_i, w_i \rangle$ where $u = 1$ or 2 denoting which user's profile term the node represents. The $path(v_i^1, v_j^2)$ denotes the set of edges between two nodes v_i^1 and v_j^2 in the semantic network. All the edges between any two nodes with different terms in the semantic network have uniform weights $\forall e \in path(v_i^1, v_j^2)$ set $wt(e) = 1$ where $wt(e)$ denotes the weight of the edge e. For any two vertices v_i^1 and v_j^2 the distance between the them is

$$len(v_i^1, v_j^2) = \begin{cases} 0, & \text{if } t_i = t_j \\ \sum_{\forall e_k \in path(v_i^1, v_j^2)} wt(e_k), & \text{otherwise} \end{cases}$$

Definition 2 (Bipartite Representation). *The bipartite graph representation G of the profiles u_1 and u_2 is a pair $G = \langle V, E \rangle$ where*

- *$V = V^1 \cup V^2$ where V^1 denotes the vertices from the first profile u_1 and V^2 denotes the vertices from the second profile u_2*
- *$V^1 = \{v_1^1, v_2^1, \ldots, v_n^1\}$ and $V^2 = \{v_1^2, v_2^2, \ldots, v_m^2\}$ where $n \leq m$ and $v_i^k = \langle t_i^k, w_i^k \rangle$ is a term.*
- *$E = \{e_{11}, e_{12}, \ldots, e_{ij}\}$ where $i = \{1, 2, \ldots, n\}$, $j = \{1, 2, \ldots, m\}$ and $len(v_i^1, v_j^2)$ denotes the path length between then vertices v_i^1 and v_j^2.*

[3] n and m are the number of terms in the first and second profile respectively.

Given the bipartite representation G, the optimal matching $E' \subseteq E$ between two vertex sets is computed using the Hungarian Algorithm [27]. The optimal bipartite graph (shown on the right side of Figure 2b) is $G' = \langle V, E' \rangle$ where $E' \subseteq E$ such that $\sum_{\forall e_{ij} \in E'} len(v_i^1, v_j^2)$ is optimal. Given the weights of vertices in the representation $W^{12} = \{w_1^1, \ldots, w_i^1, w_1^2, \ldots, w_j^2\}$, these are normalized (value [0-1]) to $W^{12'} = \{w_1^{1'}, \ldots, w_i^{1'}, w_1^{2'}, \ldots, w_j^{2'}\}$ where $\forall_{w_l^{k'} \in W^{12'}}$ is $w_l^{k'} = \frac{w_l^k}{\sum w_l^k}$.

Aggregate Path Distances: Abiding by our notion that the closer nodes with higher weights contribute more to the similarity value, we present three (slightly different) path length aggregation measures for empirical evaluation. The path distance of an edge e_{ij} in the optimal bipartite graph is defined as

$$path(e_{ij}) = \begin{cases} 1, & \text{if } len(v_i^1, v_j^2) \text{ is } 0 \\ 0, & \text{if } len(v_i^1, v_j^2) \text{ is } \infty \\ \frac{w_i^{1'} \times w_j^{2'}}{len(v_i^1, v_j^2)}, & \text{otherwise} \end{cases}$$

The Euler path distance of an edge e_{ij} in the optimal bipartite graph is defined as

$$eupath(e_{ij}) = \begin{cases} 1, & \text{if } len(v_i^1, v_j^2) \text{ is } 0 \\ 0, & \text{if } len(v_i^1, v_j^2) \text{ is } \infty \\ \frac{w_i^{1'} \times w_j^{2'}}{e^{len(v_i^1, v_j^2)}}, & \text{otherwise} \end{cases}$$

The Euler half path distance of an edge e_{ij} in the optimal bipartite graph is defined as

$$euhalf(e_{ij}) = \begin{cases} 1, & \text{if } len(v_i^1, v_j^2) \text{ is } 0 \\ 0, & \text{if } len(v_i^1, v_j^2) \text{ is } \infty \\ \frac{w_i^{1'} \times w_j^{2'}}{e^{\left(\frac{len(v_i^1, v_j^2)}{2}\right)}}, & \text{otherwise} \end{cases}$$

The aggregate distance of all the matching edges of the bipartite graph is given by the sum of their path distances.

Similarity Measures: Given two user profiles u_1 and u_2, the similarity between them using aggregate path distances in the optimal bipartite graph are defined as follows.

$$(3) \quad sim_{path}(u_1, u_2) = \frac{\sum_{\forall e_{ij} \in E'} path(e_{ij})}{min(size(terms(u_1)), size(terms(u_2))) \times max(path(e_{ij}))}$$

$$(4) \quad sim_{eupath}(u_1, u_2) = \frac{\sum_{\forall e_{ij} \in E'} eupath(e_{ij})}{min(size(terms(u_1)), size(terms(u_2))) \times max(eupath(e_{ij}))}$$

$$(5) \quad sim_{euhalf}(u_1, u_2) = \frac{\sum_{\forall e_{ij} \in E'} euhalf(e_{ij})}{min(size(terms(u_1)), size(terms(u_2))) \times max(euhalf(e_{ij}))}$$

5.4 Compound Similarity Measures

While the term vector similarity technique considers only intersecting terms while computing similarity, when two profiles actually intersect this measure is quite accurate. Therefore, we propose compound similarity measures where the similarity between intersecting profile terms are computed using cosine similarity (Equation 1), and the

similarity between the remaining profile terms are computed using our bipartite graph approaches (Equations 3, 4, and 5). More details follow.

Given two user profiles u_1 and u_2, the intersecting profile parts are denoted as u'_1 and u'_2 such that $terms(u'_1) = terms(u'_2) = terms(u_1) \cap terms(u_2)$. The remaining non-overlapping profile parts are denoted as $\hat{u_1}$ and $\hat{u_2}$ such that $terms(\hat{u_1}) = terms(u_1) \setminus terms(u_2)$ and $terms(\hat{u_2}) = terms(u_2) \setminus terms(u_1)$. The combined size of the two profiles is denoted as $N = |terms(u_1)| + |terms(u_2)|$. The size of the intersecting profile parts is $N' = |terms(u'_1)| + |terms(u'_2)|$. The size of the non-overlapping profile parts is $\hat{N} = |terms(\hat{u_1})| + |terms(\hat{u_2})|$.

The compound similarity measure based on sim_{path} (Equation 3) is

$$sim^{C}_{path} = \frac{sim_{cos}(u'_1, u'_2) \times N' + sim_{path}(\hat{u_1}, \hat{u_2}) \times \hat{N}}{N} \tag{6}$$

The compound similarity measure based on sim_{eupath} (Equation 4) is

$$sim^{C}_{eupath} = \frac{sim_{cos}(u'_1, u'_2) \times N' + sim_{eupath}(\hat{u_1}, \hat{u_2}) \times \hat{N}}{N} \tag{7}$$

The compound similarity measure based on sim_{euhalf} (Equation 5) is

$$sim^{C}_{euhalf} = \frac{sim_{cos}(u'_1, u'_2) \times N' + sim_{euhalf}(\hat{u_1}, \hat{u_2}) \times \hat{N}}{N} \tag{8}$$

6 Evaluation and Results

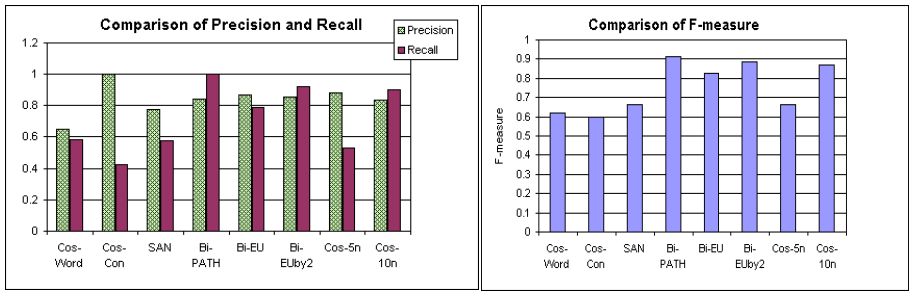
We evaluate the different algorithms described in the previous section in the context of expert finding. We use an inhouse-built software called Profile Builder to generate expert profiles using techniques described in [7] to create profiles by analysing the documents (such as web pages visited by the expert). Both the BOW (word profiles) and BOC (terms are Wikipedia concepts; Wiki profiles) representations of the experts are generated by the profile builder software. An expert finding query is correspondingly in the form of either a BOW or a BOC. For a given query profile, matching expert profiles are determined by computing similarity between the expert profile and the query profile.

Measure	**Description**
COS-Word	Cosine similarity measure between expert and query BOW profiles (Equation 1)
COS-Con	Cosine similarity measure between expert and query BOC profiles (Equation 1)
COS-5n	Mean cosine similarity between BOC profiles after 5 iterations of set spreading
COS-10n	Mean cosine similarity between BOC profiles after 10 iterations of set spreading
Bi-PATH	Compound similarity measure after graph spreading as defined in Equation 6
Bi-EU	Compound similarity measure after graph spreading as defined in Equation 7
Bi-EUby2	Compound similarity measure after graph spreading as defined in Equation 8
SAN	Similarity measure after graph spreading as defined in Equation 2

Table 1: Glossary of the Similarity Measures

A pilot study conducted as a part of the evaluation process interviewed 10 participants with expertise in different fields of computer science research. From each of the participants, 5 to 10 documents that in the participant's opinion best describe their research were collected. Along with the documents, the participants were asked to give 5 keywords for each of their document that in their opinion best described the document. Since these keywords somewhat described the expertise of the participants, they were used by the participants to provide two similarity judgments. We believe this approach

reduces the subjectivity in judging similarity and gives us more realistic values for comparison. Every participant was asked to judge the similarity between their profile and other profiles. Additionally, each of the participants judged the similarity between every pair of profiles (third person view). The mean of the subjective judgments provided by the participants were used as the base/reality values to evaluate our similarity measures. The comparison of the computed similarity value with the reality values were actually made across all user pairs. However, for evaluating the algorithms in the context of expert finding, we consider a user q to represent the query profile and evaluate similarity results of user pairs (q, x) where x is every other user (experts).



(a) Precision and Recall Observations (b) F-measure Observations

Fig. 3: Effectiveness of Similarity Measures for Expert Search (Threshold-based Match)

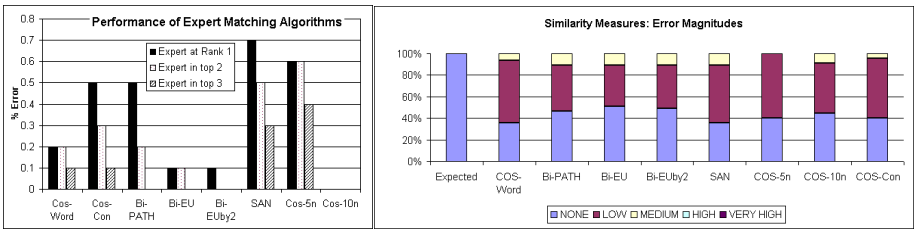
We first evaluate the effectiveness of our similarity measures in the context of short-listing a group of experts (eg: for recruitment interview). Here the selected expert profiles are those that exceed a pre-determined similarity threshold. We repeat the search for 10 query profiles over the derived expert profiles using all the approaches listed in Table 1. Figure 3 shows the results from our candidate search process where we measured precision, recall, and F-measure[4] of all the approaches. In short, precision represents the fraction of the correctly determined experts from those selected by our algorithms (based on how many of the matched results are the experts we wanted to get). Recall represents the effectiveness of the algorithm to find all experts (based on how many experts did we miss). F-measure is a compound measure, basically a harmonic mean of precision and recall. Higher values are better. From Figure 3, we are able to make the following observations.

- All the metrics based on bipartite graph mapping (Bi-*) work very well over the standard cosine similarity measurement techniques (Cos-Word and Cos-Con).
- The accuracy of set-based measures increases with the increase in the number of spreading iterations (Cos-10n performs much better than Cos-5n).
- The precision of all our approaches are almost equal while the recall varies.
- Our algorithms show significant improvements in recall when compared with the standard approaches. Our approaches Bi-* and Cos-10n exhibit upto 20% improvement.
- The recall of our Bi-PATH approach is 100% while Bi-EUby2 and Cos-10n approaches exhibit around 90% recall.
- Spreading with 5 iterations (Cos-5n) is almost equal performance to other path-based/reachability conditions for termination in a general semantic search approach (SAN). This may be suggestive of the maximum diameter of the relevant subgraph consisting of the user's concepts.

[4] We use the standard definitions of Precision, Recall and F-measure as defined in [8]

– The precision of the cosine similarity approach considering semantic concepts (Cos-Con) is 100% however it has the poorest recall of around 40%. It shows only right experts but may miss 60% of other experts.

We conclude that user profile matching through use of ontologies increases the accuracy of expert finding process and bipartite based compound measures Bi-PATH and Bi-EUby2 matches performs the best.



(a) Errors in Top 3 Selected Experts (b) Error Magnitudes over all User Pairs

Fig. 4: Accuracy of Expert Search using Different Algorithms

We next analyse the accuracy of our approaches in the context of determining an expert within the top three selections returned by our expert finding process. Here, we choose the top 3 experts based on reality values and compare those with the top 3 matches using our computed similarity metrics. The error percentage of all the approaches in this scenario is presented in Figure 4a - lower the better. As seen, our bipartite-graph based algorithms can accurately spot an expert just within its top three ranks. The Cos-Word approach has a 20% chance that the first expert returned is not the required expert. Among the top three ranks, Cos-Word still does not guarantee that a matching expert will be found because there is a 10% chance that the top three results are false positives.The set-based measures Cos-10n is the best among all the approaches with the high possibility that all the top three ranks are positive expert matches.

In order to check the effectiveness of the algorithms as a similarity measure for matching any two users, we show the magnitude of error across all the 100 user pairs. Analysis of the error magnitudes[5], as shown in Figure 4b, that our spreading based computations yield more accurate similarity judgements than the simple vector based counterparts as our bipartite approaches have the maximum number of *no errors* as a generic matching of two user profiles.

7 Conclusion

We presented a number of similarity computation measures that improve the expert finding process by accurately matching expert profiles for a query. Our approach utilises spreading as a means to capture the semantics of the terms in user profiles. The evaluation of the similarity measures shows the improvements in accuracy that is achieved over existing traditional similarity computation methods. Our bipartite graph based measures out perform all other algorithms for the specific use case of expert finding. We plan to explore use of more sophisticated techniques [24] to measure similarity at single concept level and study their effects on the profile matching. Additionally, we would like to

[5] Difference in slabs, for example expected = VERY HIGH, observed = VERY LOW results in VERY HIGH error magnitude

extend the approaches to automatically use other domain ontologies (not just Wordnet or Wikipedia) from a ontology repository like Swoogle.

References

1. Bogers, T., Kox, K., van den Bosch, A.: Using Citation Analysis for Finding Experts in Workgroups. In: Proc. DIR. (2008)
2. Demartini, G.: Finding Experts Using Wikipedia. In: FEWS. (2007)
3. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proc. IJCAI. (2007)
4. Nagarajan, M et al.: Altering document term vectors for classification: ontologies as expectations of co-occurrence. In: WWW. (2007)
5. Cantador, I et al.: A multi-purpose ontology-based approach for personalised content filtering and retrieval. In: Advances in Semantic Media Adaptation and Personalization. (2008)
6. Jung, H., Lee, M., Kang, I.S., Lee, S., Sung, W.K.: Finding topic-centric identified experts based on full text analysis. In: FEWS. (2007)
7. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using Wikipedia. In: SIGIR. (2007)
8. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
9. Dhyani, D., Ng, W.K., Bhowmick, S.S.: A survey of web metrics. ACM Comput. Surv. **34**(4) (2002)
10. van Rijsbergen, C.J.: Information Retrieval. Butterworth (1979)
11. Hamers, L et al.: Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Inf. Process. Manage. **25**(3) (1989)
12. Biswas, H., Hasan, M.: Using Publications and Domain Knowledge to Build Research Profile: An Application in Automatic Reviewer Assignment. In: Proc. ICICT. (2007)
13. : TREC Enterprise Track. http://www.ins.cwi.nl/projects/trec-ent/wiki/ (2005)
14. Castells, P., Fernández, M., Vallet, D., Mylonas, P., Avrithis, Y.S.: Self-tuning personalized information retrieval in an ontology-based framework. In: OTM Workshops. (2005)
15. Tsatsaronis, G., Vazirgiannis, M., Androutsopoulos, I.: Word sense disambiguation with spreading activation networks generated from thesauri. In: Proc. IJCAI. (2007)
16. Véronis, J., Ide, N.: Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. In: In Proc. COLING. (1990)
17. Mao, M.: Ontology mapping: An information retrieval and interactive activation network based approach. In: Proc. ISWC. (2007)
18. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet:: Similarity - Measuring the Relatedness of Concepts. In: Proc. AAAI. (2004)
19. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proc. IJCAI. (1995)
20. Zhu, H., Zhong, J., Li, J., Yu, Y.: An Approach for Semantic Search by Matching RDF Graphs. In: FLAIRS. (2002)
21. Li, Y., Bandar, Z., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng. **15**(4) (2003)
22. Schickel-Zuber, V., Faltings, B.: Oss: A semantic similarity function based on hierarchical ontologies. In: Proc. IJCAI. (2007)
23. Iosif, E., Potamianos, A.: Unsupervised semantic similarity computation using web search engines. In: Proc. of WI. (2007)
24. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics **32**(1) (2006)
25. Crestani, F.: Application of Spreading Activation Techniques in Information Retrieval. Artif. Intell. Rev. **11**(6) (1997) 453–482
26. Thiagarajan, R., Manjunath, G., Stumptner, M.: Computing semantic similarity using ontologies. HP Labs Technical Report HPL-2008-87 (2008)
27. Kuhn, H.W.: The Hungarian Method for the Assignment Problem. Naval Research Logistic Quarterly **2** (1955) 83–97

Finding Experts on the Semantic Desktop

Gianluca Demartini and Claudia Niederée

L3S Research Center
Leibniz Universität Hannover
Appelstrasse 9a, 30167 Hannover, Germany
{demartini,niederee}@L3S.de

Abstract. Expert retrieval has attracted deep attention because of the huge economical impact it can have on enterprises. The classical dataset on which to perform this task is company intranet (i.e., personal pages, e-mails, documents). We propose a new system for finding experts in the user's desktop content. Looking at private documents and e-mails of the user, the system builds expert profiles for all the people named in the desktop. This allows the search system to focus on the user's topics of interest thus generating satisfactory results on topics well represented on the desktop. We show, with an artificial test collection, how the desktop content is appropriate for finding experts on the topic the user is interested in.

1 Introduction

Finding people who are expert on certain topics is a search task which has been mainly investigated in the enterprise context. Especially in big enterprises, topic areas can range very much also because of diverse and distributed data sources. This peculiarity of enterprise datasets can highly affect the quality of the results of the expert finding task [15, 16].

It is important to provide the enterprise managers with high quality expert recommendation. The managers need to build new project teams and to find people who can solve problems. Therefore, a high-precision tool for finding experts is needed. Moreover, not only managers need to find experts. In a highly collaborative environment where the willingness of sharing and helping other team members is present, all the employees should be able to find out to which colleague to ask for help in solving issues.

If we want to achieve high-quality results while searching for experts, considering the user's desktop content makes the search much more focused on the user's interests also because the desktop dataset will contain much more expertise evidence (on such topics) than the rest of the public enterprise intranet. Classic expert search systems [9, 30, 21, 25, 26, 17] work on the entire enterprise knowledge available. This means that they use shared repository, e-mails history, forums, wikis, databases, personal home pages, and all the data that an enterprise creates and stores. This makes the system to consider a huge variety of topics, for example, from accountability to IT specific issues. Our solution

focuses on using the user's desktop content as expertise evidence allowing the system to focus on the user's topics of interest thus providing high quality results for queries about such topics.

The system we propose is first indexing the desktop content also using metadata annotation that are produced by the Social Semantic Desktop system Nepomuk [19]. Our expert search system creates a vector space that includes the documents and the people that are present in the desktop content. After this step, when the desktop user issues a query of the type "*Find experts on the topic...*"+*keywords* the system shows a ranked list of people that the user can contact for getting help. Preliminary experiments show the high precision of the expert search results on topics which are covered by the desktop content. A limitation of our system is that it can return only people that are present on the user's desktop. Therefore, the performances are poor when the desktop content (i.e., number of items and people) is limited, as for example for new employees, or when the queries are different from the main topics represented in the desktop. The main contributions of the paper are:

- the description of how the Beagle++ system creates metadata regarding documents and people (Section 2.1).
- a new system for finding experts on a semantic desktop (Section 2.2).
- the description of possible test datasets: one composed of fictitious data and one containing real desktop content (Section 3).
- preliminary experimental results showing how a focused dataset leads to high-quality expert search results (Section 4).
- a review of the previous systems and formal models presented in the field of expert search and Personal Information Management (PIM) (Section 5).

2 System Architecture

2.1 Generating Metadata about People

In order to identify possible expert candidates and link them to desktop items, we used extractors from the Beagle++ Dekstop Search Engine[1] [2] [13, 8]. These extractors identify documents and e-mails authors by analysing the structure and the content of each file. For storing the produced metadata (see Figure 1) we employ the RDF repository developed in the Nepomuk project [19] based on Sesame[3] for storing, querying, and reasoning about RDF and RDF Schema, as well as on Lucene[4], which is integrated with the Sesame framework via the LuceneSail [27], for full-text search.

An additional step is the entity linkage applied to the identified candidates. For example, a person in e-mails is described by an e-mail address, whereas in a publication by the author's name. Other causes for the appearance of different

[1] `http://beagle2.kbs.uni-hannover.de`
[2] `http://www.youtube.com/watch?v=Ui4GDkcR7-U`
[3] `http://www.openrdf.org`
[4] `http://lucene.apache.org`

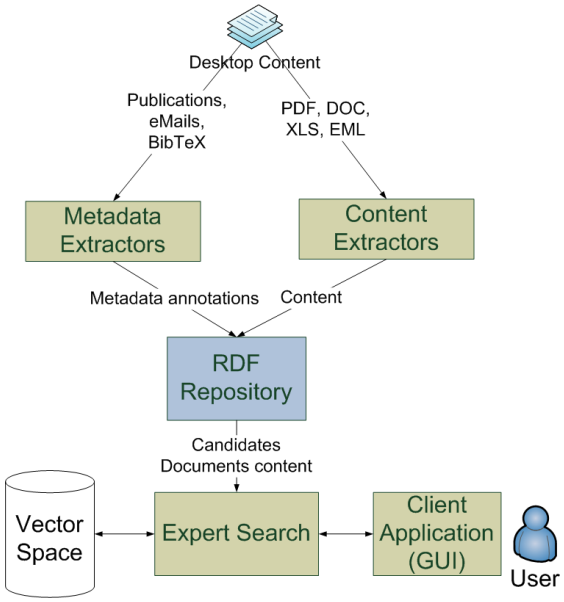


Fig. 1. An overview of how the desktop content is extracted and given in input to the expert search component for indexing. A client application is providing a user interface to the expert search service.

references to the same entity are misspellings, the use of abbreviations, initials, or the actual change of the entity over time (e.g., the e-mail address of a person might change). Again, we exploit a component of the Beagle++ search system for producing information about the linkage.

At this point, we obtained a repository describing desktop items content and metadata. In the next section we explain how we can exploit this data and metadata for finding experts in the semantic desktop content.

2.2 Leveraging Metadata for People Search

In the Nepomuk system, the service of Expert Recommendation[5] aims at providing the user with a list of experts (i.e., people) on a given topic. The experts are selected among a list of persons referral to in the desktop. In order to do so, the component needs to extract, out of the RDF repository, some information about the content of documents and e-mails and also a list of expert candidates (see Figure 1).

Thanks to the Beagle++ system, relations between people and documents are identified and stored in the repository. Entity Linkage identifies references pointing to the same entity by gathering clues as, for example, a person in e-mails described by an e-mail address, whereas in a publication by the author's name. In Beagle++, searching using a person's surname retrieves publications

[5] `http://dev.nepomuk.semanticdesktop.org/wiki/ExpertRecommender`

in which her surname appears as part of an author field as well as e-mails in which her e-mail address appears as part of the sender or receiver fields. This is obtained linking together the objects that refer to the same real world entities [20].

The expert search system we propose can leverage on the extracted relations between documents and people as well as on the linkage information between different representations (e.g., surname and e-mail address). The first step is to create an inverted index for documents: a vector representation of each publication, e-mail, and text-based resources on the desktop is created. Then, for each expert candidate referral to in the desktop, her position into the vector space is computed by linear combination of the resources related with her, using the relation strength as weight. At this point, each candidate expert is placed into the space and a query vector, together with a similarity measure (e.g., cosine similarity), can be used to retrieve a ranked list of experts. The fact that documents are indexed before candidates implies that the dimensions of the vector space are defined by the set of terms present in the desktop collection. This means that the topics of expertise that represents the candidates are those inferred from the documents.

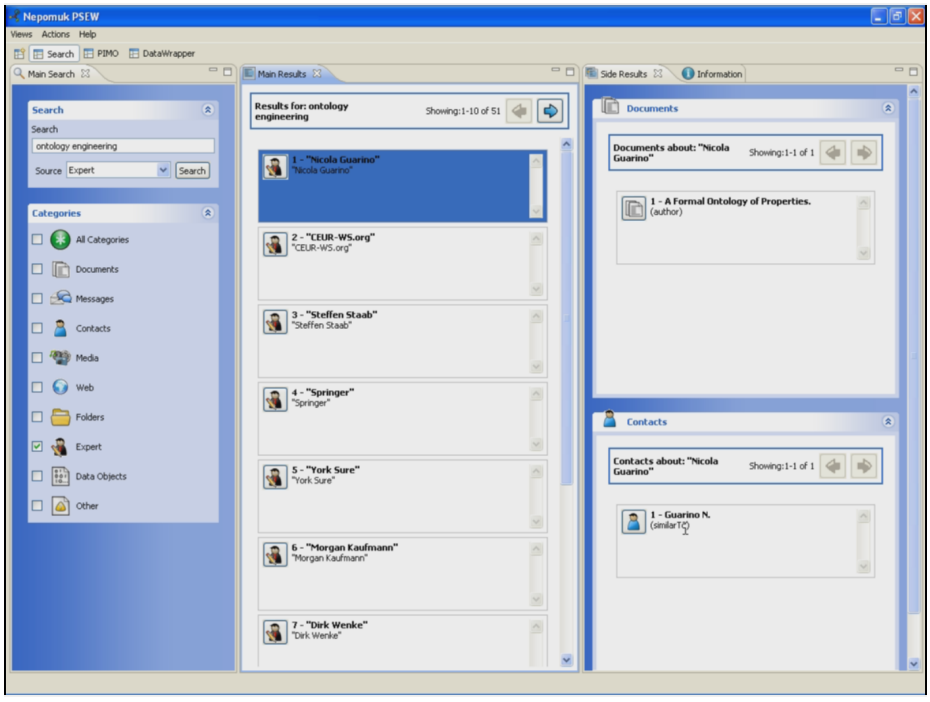


Fig. 2. A client application for searching experts on the semantic desktop.

A client application can then use the Nepomuk Expert Recommendation service (which implements the system described in this paper) by providing a keyword query taken from the user. A screenshot of a possible client application is shown in Figure 2. In the top-left corner the user can provide a keyword query and the choice of looking for experts. In the central panel a ranked list of people is presented as result of the query. In the right pane, resources related to the selected expert are shown.

3 Desktop Search Evaluation Datasets

Evaluation of desktop search algorithms effectiveness is a difficult task because of the lack of standard test collections. The main problem of building such test collection is the privacy concerns that data providers might have while sharing personal data. The privacy issue is major as it impedes the diffusion of personal desktop data among researches. Some solutions for overcoming these problems have been presented in previous work [11, 12].

In this section we describe two possible datasets for evaluating the effectiveness of finding experts using desktop content as evidence of expertise. One is a fictitious desktop dataset representing two hypothetical personas. This dataset has been manually created in the context of the Nepomuk project with the goal of providing a publicly available desktop dataset with no privacy concerns. As at present, the access to the actual data is still restricted. The second one is a set of real desktop data provided by 14 employees of a research center.

3.1 Fictitious Data

In order to obtain reproducible and comparable experimental results there is a need for a common test collection. That is, a set of resources, queries, and relevance assessments that are publicly available. In the case of PIM the privacy issue of sharing personal data has to be faced. For solving this issue the team working on the Nepomuk project has created a collection of desktop items (i.e., documents, e-mails, contacts, calendar items, ...) for some imaginary personas representing hypothetical desktop users. In this paper we describe two desktop collections built in this context.

The first persona is called Claudia Stern[6]. She is a project manager and her interests are mainly about ontologies, knowledge management, and information retrieval. Her desktop contains 56 publications about her interests, 36 e-mails, 19 Word documents about project meetings and deliverables, 12 slides presentations, 17 calendar items, 2 contacts, and an activity log collected while a travel was being arranged (i.e., flight booking, hotel reservation, search for shopping places) containing 122 actions. These resources have been indexed using the Beagle++ system obtaining a total of 22588 RDF triples which have been stored in the RDF repository.

The second persona is called Dirk Hagemann[7]. He works for the project that Claudia manages and his interests are similar to those of Claudia . His desktop

[6] `http://dev.nepomuk.semanticdesktop.org/wiki/Claudia`
[7] `http://dev.nepomuk.semanticdesktop.org/wiki/Dirk`

contains 42 publications, 9 e-mails, 19 Word documents, and 7 text files. These resources have been indexed using the Beagle++ system obtaining a total of 11914 RDF triples which have been stored in the RDF repository.

3.2 Real Data

For evaluating the retrieval effectiveness of a personal information retrieval system, a test collection that accurately represents the desktop characteristics is needed. However, given highly personal data that users usually have on their desktops, currently there are no desktop data collections publicly available. Therefore, we created for experimental purposes our internal desktop data collection. More detail can be found in [11].

The collection that we created - and which are currently using for evaluation experiments - is composed of data gathered from the PCs of 14 different users. The participant pool consists of PhD students, PostDocs and Professors in our research group. The data has been collected from the desktop contents present on the users' PCs in November 2006.

Privacy Preservation In order to face the privacy issues related to providing our personal data to other people, a written agreement has been signed by each of the 14 providers of data, metadata and activities. The document is written with implication that every data contributor is also a possible experimenter. The text is reported in the following:

> **L3S Desktop Data Collection**
> **Privacy Guarantees**
>
> – I will not redistribute the data you provided me to people outside L3S. Anybody from L3S whom I give access to the data will be required to sign this privacy statement.
> – The data you provided me will be automatically processed. I will not look at it manually (e.g. reading the e-mails from a specific person). During the experiment, if I want to look at one specific data item or a group of files/data items, I will ask permission to the owner of the data to look at it. In this context, if I discover possibly sensitive data items, I will remove them from the collection.
> – Permissions of all files and directories will be set such that only the *l3s-experiments-group* and the super-user has access to these files, and that all those will be required to sign this privacy statement.

Currently Available Data The desktop items that we gathered from our 14 colleagues, include e-mails (sent and received), publications (saved from e-mail attachments, saved from the Web, authored / co-authored), address books and calendar appointments. A distribution of the desktop items collected from each user can be seen in Table 1:

User#	E-mails	Publications	Addressbooks	Calendars
1	109	0	1	0
2	12456	0	0	0
3	4532	1054	1	1
4	834	237	0	0
5	3890	261	1	0
6	2013	112	0	0
7	218	28	0	0
8	222	95	1	0
9	0	274	1	1
10	1035	31	1	0
11	1116	157	1	0
12	1767	2799	0	0
13	1168	686	0	0
14	49	452	0	0
Total	29409	6186	7	2
Avg	2101	442	0.5	0.1

Table 1. Resource distribution over the users in the L3S Desktop Data Collection.

A total number of 48,068 desktop items (some of the users provided a dump of their desktop data, including all kinds of documents, not just e-mails, publications, address books or calendars) has been collected, representing 8.1GB of data. On average, each user provided 3,433 items.

In order to emulate a standard test collection, all participants provided a set of queries that reflects typical activities they would perform on their desktop. In addition, each user was asked to contribute their activity logs, related to the period until the point at which the data were provided. All participants defined their *own* queries, related to their activities, and performed search over the reduced images of their desktops, as mentioned above.

4 Preliminary Experiments

We used the Dirk and Claudia datasets (see Section 3.1) in order to perform some initial evaluation of our system for finding experts. We created some queries that match the personas interests imagining which kind of experts they would need to find.

The expert search queries on the Dirk's desktop are:

- ontology engineering
- pagerank
- religion

The expert search queries on the Claudia's desktop are:

- ontology engineering
- ranking in information retrieval
- document search

We issued the same query (i.e., "ontology engineering") on the two datasets in order to compare the results. Dirk and Claudia have the same interest for ontologies but the Dirk desktop contains less data than Claudia's. Table 2 shows the top 5 results on the two datasets. We can see that the results are similar as Dirk and Claudia also share some publications on their desktops. While all the top 5 retrieved people have been working on the topic, the ranking might be improved. For example, the candidate "Dirk Wenke" has less experience than "Nicola Guarino" or "Rudi Studer" on the topic. The explanation of this result is that only local evidence of expertise is used. The quality might be improved by looking at evidence on the web (e.g., DBLP[8] pages).

	Dirk	Claudia
1	Steffen Staab	Steffen Staab
2	York Sure	Riichiro Mizoguchi
3	Rudi Studer	Dirk Wenke
4	Dirk Wenke	York Sure
5	Nicola Guarino	Rudi Studer

Table 2. Top 5 results for the query "ontology engineering".

On Dirk's data we issues the query "pagerank" meaning the famous link based algorithm proposed by Brin and Page in [7]. The top 5 results are presented in Table 3. We can see, again, that all the retrieved candidates have some experience on the topic, but the ordering is not good enough. The authors of the algorithm are placed fourth and fifth while they should be at the top of the list. The first three retrieved candidates have been working on the P2P version of the algorithm.

	Dirk	Claudia	Claudia
	pagerank	ranking in information retrieval	document search
1	Karthikeyan Sankaralingam	Sergey Brin	Jon Kleinberg
2	Simha Sethumadhavan	Karl Aberer	Karl Aberer
3	James C. Browne	Lawrence Page	Eli Upfal
4	Sergey Brin	Jon Kleinberg	Sergey Brin
5	Lawrence Page	Eli Upfal	Monika Henzinger

Table 3. Top 5 results for the query "pagerank" on Dirk's desktop. Top 5 results for the queries "ranking in information retrieval" and "document search" on Claudia's desktop.

The query "religion" on Dirk's desktop, as expected, returned no results. This can be explained because there is no evidence of expertise on such topic in this dataset.

[8] `http://www.informatik.uni-trier.de/~ley/db/`

Finally, we discuss the last two queries on Claudia's dataset. We created queries on very similar topics (i.e., "ranking in information retrieval" and "document search") in order to compare the results. The results are shown in Table 3. We can see that the top 5 results are similar but the ranking. In this case it is hard to say which the best ranking should be as all the retrieved candidates have strong experience on the topic and deciding who is the most expert is highly subjective.

In conclusion, we have seen that the effectiveness of finding experts using the desktop content highly depends on the available resources. If the user queries for experts on topics well represented on her desktop, then the results can be satisfactory. If the query is off-topic then the results can be poor or even be missing. Moreover, further improvements are needed on the ranking function used. A novel measure replacing the cosine similarity used in this experiments might be used.

5 Discussion of Related Work

In this section we describe and discuss the previous work in the field of Expert Search and PIM. We show how existing systems have been designed, which formal models have been proposed, which PIM systems can be extended with expert search functionalities.

5.1 Expert Search Systems

Several expert search systems have been proposed in the last years. These systems use different information sources and features like social network information [9], co-occurrences of terms and changes in the competencies of people over time [30], rule-based models and FOAF[9] data [21]. For the web, a different context from the enterprise search one, one of the approaches proposed [29] focuses on scenarios like Java Online Communities where experts help newcomers or collaborate with each other, and investigated several algorithms that build on answer-reply interaction patterns, using PageRank and HITS authority models as well as additional algorithms exploiting link information in this context. We are not aware of any system for finding experts on the desktop.

The Enterprise PeopleFinder [25, 26] also known as P@noptic Expert [17] first builds a candidate profile attaching all documents related to that candidate in one big document giving different weights to the documents based on their type.

An interesting distinction has been made between *expert finding* and *expert profiling* in [4]. The former approach aims at first retrieving the documents relevant to the query and then extract the experts from them. The latter first builds a profile for each candidate and then matches the query with the profiles without considering the documents anymore [5].

[9] `http://www.foaf-project.org`

5.2 Expert Search Models

All systems mentioned up to now use different ad-hoc techniques but do not formally define retrieval models for experts. Some first steps in this direction have been made: probabilistic models [18] and language models [1–3] have been proposed. Another model for expert search proposed in [23] views this task as a voting problem. The documents associated to a candidate are viewed as votes for this candidate's expertise. In [24] the same authors extended the model including relevance feedback techniques, which is an orthogonal issue. More recently, focus has been put on finding high quality relationships between documents and people and evidence of expertise [28, 22, 6].

5.3 Personal Information Management Systems

A lot of research have been also done in the field of PIM. The most relevant area is the one of desktop search. Finding items on the desktop is not the same task as finding documents on the web. Several commercial systems have been proposed (e.g., Google, Yahoo!, Microsoft). Our expert finding system builds on top of the Beagle++ system: a semantic desktop search engine [8]. Beagle++ exploits the implicit semantic information residing at the desktop level in order to enhance desktop search. Moreover, it creates metadata annotations, thanks to its extractors, that can be reused by our expert finding system.

One important issue in the field of PIM is the evaluation of retrieval effectiveness. Retrieval systems are usually evaluated using standard testbeds (e.g., TREC[10]). In PIM such testbeds are not available mainly because of the privacy issues of sharing personal data. A way to overcome this problem is to create small collections internally to each research group [11].

The Nepomuk project aims at developing a framework for the Social Semantic Desktop. Our expert finding system is integrated in the Nepomuk system providing the user of the semantic desktop this additional search functionality.

If we want to find experts on the desktop, then a crucial task is to extract people names out of full text. Many techniques have been proposed and can be reused for this step. Possible solutions to the problem of measuring similarity between two named entities are presented in [14], how to pre-process a document collection in order to extract names from documents such as e-mail has been proposed in [10].

6 Conclusions and Future Work

In this paper we presented a system for finding experts on the semantic desktop. The approach works as follow. The desktop content is first indexed: metadata is extracted and an RDF repository is built with information about persons and documents. Then, a vector space containing candidate experts and documents is created by exploiting the relations existing between them. Once the documents as well as the candidates are placed into the vector space, a query vector can be placed into the space and a ranked list of experts can be obtained using a

[10] `http://trec.nist.gov`

similarity measure. We used two artificial datasets for performing preliminary experiments. The results show that search results are good for topics that are well represented in the desktop content and poor for others. Effectiveness might be improved by exploiting external evidence of expertise as, for example, web pages. The Beagle++ system indexes visited web pages and, therefore, it could include information from the web also leveraging on semantic technologies such as microforomats or RDFa. Moreover, evidence of expertise contained in both the enterprise intranet and the desktop could be combined in order to generate better results. As future work, we aim at performing a user study using the collection made of data from real user desktops (see Section 3.2) with the goal of evaluating the effectiveness of the expert finding system presented in this paper.

Acknowledgements. We thank the anonymous reviewers for their valuable comments. This work is partially supported by the EU Large-scale Integrating Project OKKAM[11] - Enabling a Web of Entities (contract no. ICT-215032) and by the NEPOMUK[12] project funded by the European Commission under the 6th Framework Programme (IST Contract No. 027705).

References

1. L. Azzopardi, K. Balog, and M. de Rijke. Language modeling approaches for enterprise tasks. *The Fourteenth Text REtrieval Conference (TREC 2005)*, 2006.
2. K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. *Proceedings of the 29th SIGIR conference*, pages 43–50, 2006.
3. K. Balog and M. de Rijke. Finding experts and their Details in e-mail corpora. *Proceedings of the 15th international conference on World Wide Web*, pages 1035–1036, 2006.
4. K. Balog and M. de Rijke. Searching for people in the personal work space. *International Workshop on Intelligent Information Access (IIIA-2006)*, 2006.
5. K. Balog and M. de Rijke. Determining Expert Profiles (With an Application to Expert Finding). *Proceedings of IJCAI-2007*, pages 2657–2662, 2007.
6. K. Balog and M. de Rijke. Associating people and documents. In *ECIR*, pages 296–308, 2008.
7. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.
8. I. Brunkhorst, P. A. Chirita, S. Costache, J. Gaugaz, E. Ioannou, T. Iofciu, E. Minack, W. Nejdl, and R. Paiu. The beagle++ toolbox: Towards an extendable desktop search architecture. *Proceedings of the Semantic Desktop and Social Semantic Collaboration Workshop (SemDesk 2006)*, November 2006.
9. C. Campbell, P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. *Proceedings of the 12th ACM Conference on Information and Knowledge Management (CIKM'03)*, pages 528–531, 2003.
10. V. Carvalho and W. Cohen. Learning to Extract Signature and Reply Lines from Email. *Proceedings of the Conference on Email and Anti-Spam*, 2004.
11. S. Chernov, G. Demartini, E. Herder, M. Kopycki, and W. Nejdl. Evaluating personal information management using an activity logs enriched desktop dataset. In *Proceedings of 3rd Personal Information Management Workshop (PIM 2008)*, 2008.

[11] `http://fp7.okkam.org`
[12] `http://www.nepomuk.semanticdesktop.org`

12. S. Chernov, P. Serdyukov, P.-A. Chirita, G. Demartini, and W. Nejdl. Building a desktop search test-bed. In *ECIR '07: Proceedings of the 29th European Conference on Information Retrieval*, pages 686–690, 2007.
13. P.-A. Chirita, S. Costache, W. Nejdl, and R. Paiu. Beagle^{++}: Semantically enhanced searching and ranking on the desktop. In *ESWC*, pages 348–362, 2006.
14. W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string distance metrics for name-matching tasks. *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, 2003.
15. N. Craswell, A. de Vries, and I. Soboroff. Overview of the TREC 2005 Enterprise Track.
16. N. Craswell, A. de Vries, and I. Soboroff. Overview of the TREC-2006 Enterprise Track.
17. N. Craswell, D. Hawking, A. Vercoustre, and P. Wilkins. P@noptic Expert: Searching for Experts not just for Documents. *Ausweb*, 2001.
18. H. Fang and C. Zhai. Probabilistic Models for Expert Finding. *Proceedings of 29th European Conference on Information Retrieval (ECIR'07)*, pages 418–430, 2007.
19. T. Groza, S. Handschuh, K. Moller, G. Grimnes, L. Sauermann, E. Minack, M. Jazayeri, C. Mesnage, G. Reif, and R. Gudjonsdottir. The NEPOMUK Project-On the way to the Social Semantic Desktop. *Proceedings of I-Semantics'07*, pages 201–211.
20. E. Ioannou, C. Niederée, and W. Nejdl. Probabilistic entity linkage for heterogeneous information spaces. In *CAiSE*, 2008.
21. J. Li, H. Boley, V. C. Bhavsar, and J. Mei. Expert finding for eCollaboration using FOAF with RuleML rules. *Montreal Conference on eTechnologies (MCTECH)*, 2006.
22. C. Macdonald, D. Hannah, and I. Ounis. High quality expertise evidence for expert search. In *ECIR*, pages 283–295, 2008.
23. C. Macdonald and I. Ounis. Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task. *Proceedings of the 15th ACM Conference on Information and Knowledge Management (CIKM'06)*, pages 387–396, 2006.
24. C. Macdonald and I. Ounis. Using Relevance Feedback in Expert Search. *Proceedings of 29th European Conference on Information Retrieval (ECIR'07)*, pages 431–443, 2007.
25. A. McLean, A. Vercoustre, and M. Wu. Enterprise PeopleFinder: Combining Evidence from Web Pages and Corporate Data. *Proceedings of Australian Document Computing Symposium*, 2003.
26. A. McLean, M. Wu, and A. Vercoustre. Combining Structured Corporate Data and Document Content to Improve Expertise Finding. *Arxiv preprint cs/0509005*, 2005.
27. E. Minack, L. Sauermann, G. Grimnes, C. Fluit, and J. Broekstra. The Sesame LuceneSail: RDF Queries with Full-text Search. Technical report, NEPOMUK 2008-1, February 2008.
28. P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. In *ECIR*, pages 309–320, 2008.
29. J. Zhang, M. Ackerman, and L. Adamic. Expertise Networks in Online Communities: Structure and Algorithms. *Proceedings of the 16th international conference on World Wide Web*, pages 221–230, 2007.
30. J. Zhu, A. Gonçalves, V. Uren, E. Motta, and R. Pacheco. Mining Web Data for Competency Management. *Web Intelligence '05*, pages 94–100, 2005.

Requirements for expertise location systems in biomedical science and the Semantic Web

Titus Schleyer[1], Heiko Spallek[1], Brian S. Butler[2], Sushmita Subramanian[3], Daniel Weiss[4], M. Louisa Poythress[5], Phijarana Rattanathikum[6], Gregory Mueller[7]

[1]School of Dental Medicine and [2]Joseph M Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA
{titus, hspallek, bbutler}@pitt.edu
[3]Intel Corporation, Santa Clara, CA
[4]The MITRE Corporation, Bedford, MA
[5]Brulant, Inc., Beachwood, OH
[6]Adobe Systems Incorporated, San Jose, CA
[7]DeepLocal, Inc., Pittsburgh, PA

Abstract. Recent trends in science are increasing the need for researchers to form collaborations. To date, however, electronic systems have played only a minor role in helping scientists do so. This study used a literature review, and contextual inquiries and semistructured interviews with biomedical scientists to develop a preliminary set of requirements for electronic systems designed to help optimize how biomedical researchers choose collaborators. We then reviewed the requirements in light of emerging research on expertise location using the Semantic Web. The requirements include aspects such as comprehensive, complete and up-to-date online profiles that are easy to create and maintain; the ability to exploit social networks when searching for collaborators; information to help gauge the compatibility of personalities and work styles; and recommendations for effective searching and making "non-intuitive" connections between researchers. The Semantic Web offers significant opportunities for operationalizing the requirements, for instance through aggregating profile data from disparate sources, annotating contributions to social media using methods such as Semantically Interlinked Online Communities, and concept-based querying using ontologies. Future work should validate the preliminary requirements and explore in detail how the Semantic Web can help address them.

Keywords: expertise location, requirements, Semantic Web, biomedical research

1 Introduction

Increased collaboration across all fields of biomedical science has emerged as one possible way to achieve greater success and progress in combating disease and improving health. "Team science," "networked science" and inter/multi-disciplinary research [1] are terms used to denote collaborative approaches expected to solve research problems of ever-growing complexity. Programmatic initiatives such as the

Roadmap[1] and the Clinical and Translational Science Award (CTSA)[2] programs of the National Institutes of Health (NIH) demonstrate that funding agencies and research organizations are not just passively observing this trend, but are actively encouraging it.

In the process, many academic/research institutions are extending the scale and scope of their research portfolio [2] and the numbers of their research faculty, thus making more individuals available for collaboration, either locally or remotely. As a wider range of collaborations is becoming recognized as valuable, many researchers are beginning to expand their collaborative horizons. At the same time, the Internet is making locating collaborators easier. In fact, modern communication and collaborative technologies increase the number of potential collaborators by making many remote collaborations once considered impractical feasible.

At the same time, expertise location has been, and continues to be, a significant challenge for many organizations [3,4]. Scientists often turn to colleagues or the published literature to find collaborators [5]. However, these approaches do not scale well in the context of an increasing pool of potential collaborators. As the universe of potential collaborators and information about them grows, the time and effort needed to evaluate each collaborative opportunity remains the same.

A newer method for finding collaborators is to use databases of researchers partially or exclusively designed for the purpose. Knowledge management systems of this type, which include "expertise locating systems," [6] "knowledge communities," [7] and "communities of practice," [8] all embody, to varying degrees, the ability to find experts and, by extension, potential collaborators. The CSCW literature contains numerous examples of such systems [9-12]. Most of these systems are designed to help a person solve a specific problem at a particular point in time. However, scientists seeking collaborators face a bigger challenge. Not only are they looking for the most qualified expert, but they also plan to enter into a more or less long-term relationship. Evaluating an individual's promise for such a relationship requires information, engagement and effort much beyond what is needed for finding an expert for singular (or even episodic) problem-solving. Only few reports of expertise location systems in academia have been published [11,13]. While many commercial offerings, such as the Community of Science (COS; www.cos.com), LinkedIn (www.linkedin.com), Index Copernicus Scientists (scientists.indexcopernicus.com), BiomedExperts (www.biomedexperts.com) and Research Crossroads (www.researchcrossroads.com), purport to make it easier to help scientists find collaborators, no reports in the literature describe how well these systems actually do so.

The Semantic Web is a technology with significant promise to ameliorate the expertise location problem [14]. As individuals create an increasing number of "digital trails" of their work processes and products, more information about their activities and relationships becomes computationally accessible. However, expertise location systems that leverage data from the Semantic Web must be constructed with the needs and requirements of the end user in mind. We therefore have organized this paper in two parts. We first present a set of preliminary requirements for expertise location

[1] NIH Roadmap for Medical Research, http://nihroadmap.nih.gov/
[2] Clinical and Translational Science Awards, http://www.ctsaweb.org/

systems for biomedical scientists. Second, we discuss the requirements in light of technological capabilities and challenges of the Semantic Web.

2 Methods

This study drew on several methodological approaches in order to develop a rich understanding of how scientific collaborations are established and what requirements should inform the design of expertise location systems. The methods we used included (1) affinity diagramming of issues in scientific collaboration; (2) a literature review of expertise location in computer-supported cooperative work and other disciplines; (3) contextual inquiries with 10 biomedical scientists; and (4) findings from 30 semistructured interviews with biomedical scientists from a variety of disciplines.

To develop the affinity diagram, the members of the project team (which consisted of all authors) recorded thoughts, ideas and observations regarding the establishment of scientific collaborations and then took turns arranging them into naturally-forming categories. The team then rearranged the groups to form a hierarchy that revealed the major issues of the domain. The most prominent groups were then adopted as the foci of exploratory investigations, specifically the literature search and contextual inquiries.

We searched the literature using keywords including "expertise locating systems," "expertise location systems," "expertise management systems," "knowledge communities," "knowledge management" and "knowledge management systems," "communities of practice," and "virtual communities" in the field of biomedical research, informatics, computer science and information science. The databases we searched were MEDLINE, the ISI Web of Science, the ACM Portal and the IEEE Digital Library (all available years).

Contextual inquiry (CI) [15] sessions were performed with ten researchers from a range of disciplines and levels of seniority at Carnegie Mellon University and the University of Pittsburgh. Because we could not directly observe researchers in the process of forming collaborations, we mainly focused on retrospective accounts. The contextual inquiries were complemented by findings from 30 semistructured interviews with scientists. The interviews focused on current and previous collaborations, locating collaborators, solving problems in research, and information needs and information resource use of participants. Four faculty researchers (including three authors: TKS, HS, BB) and one staff member conducted the interviews individually with a convenience sample of scientists from the six Health Science Schools at the University of Pittsburgh.

While conducting our background studies, we formulated a running list of requirements for systems that help optimize how scientists choose collaborators. We generated this list using an approach similar to grounded theory [16], in which models and hypotheses are progressively inferred from the data. We kept a record of the evidence that supported each requirement, e.g. statements of our study participants or findings from the literature, as well as of factors that would modify its validity or applicability. The studies conducted as part of this project were approved by the Univer-

sity of Pittsburgh Institutional Review Board (IRB approval numbers: 0612065 and PRO07050299).

Once the list of requirements was final, we reviewed the literature about the Semantic Web with a particular focus on expertise location. We used this literature to inform the discussion of the capabilities and challenges of the Semantic Web in light of the requirements we formulated.

3 Results

3.1 Preliminary requirements for expertise location systems in biomedical science

The following 10 requirements for expertise location systems have been ordered loosely in an attempt to group related items.

(1) The effort required to create and update an online profile should be commensurate with the perceived benefit of the system.
Many current online networking systems for scientists, such as the COS, require a significant amount of effort to create and maintain a comprehensive profile. Many scientists considered this investment of time and effort difficult to justify as there is no clear gain to being part of the system. Only a few researchers we interviewed, specifically junior ones or those new to the organization, indicated that COS and/or the Faculty Research Interests Project (FRIP) at the University of Pittsburgh [11] helped them find collaborators. Several commented that they had tried to use COS and/or FRIP, but abandoned them when their attempt at finding a collaborator through them was not successful.

(2) Online profiles should present rich and comprehensive information about potential collaborators in an organized manner to reduce the effort involved in making collaboration decisions.
The Internet makes a significant amount of information available about individual scientists, but unfortunately in a very fragmented and inhomogeneous manner. Our background research showed that at present, researchers sometimes use multiple information sources such as MEDLINE, Google Scholar, the ISI Web of Science and other databases to evaluate a potential collaborator. Retrieving, collating and reviewing information from these sources, however, often takes more time and effort than the individual is willing to expend. An expertise location system should collate and organize this information and present it to collaboration seekers in an easy-to-use format in order to reduce the effort involved in choosing collaborators.

(3) Online profiles should to be up-to-date, because some information they contain has a short lifespan.
At its core, choosing a collaborator is an attempt to predict how someone else will behave in the future. While knowledge about past behavior can be useful for doing so,

the value of this information declines with time. Out-of-date profiles reduce the usefulness of information that collaboration seekers require. On the other hand, not all information in a profile is subject to the same rate of decay. Information about professional degrees of a collaborator tends to be relatively static, while publication topics and activity may not always reflect an individual's current research focus and productivity.

(4) Researchers should be able to exploit their own and others' social networks when searching for collaborators.
Social networks have been suggested as important structures for finding expertise and information [17]. Established researchers often use existing connections with colleagues as their primary resource for locating new collaborators. Junior researchers, with few or no contacts within the desired field, may have significant difficulty initiating collaborations that way. Many scientists in our study indicated they are more likely to contact a colleague whom they think will know someone with the required expertise than cold-call a stranger. In addition, many emphasized the key role that deans, department heads and other well-connected individuals in the organization play in helping establish collaborations. The advantages of a mediated form of contact are that it may make it more likely that two parties will be compatible, increase the chances of a timely response, and provide a less intimidating method of contact.

(5) The system should model proximity, which influences the potential success of collaboration in several respects.
Physical proximity, social proximity, organizational proximity, and proximity in terms of shared research interests are all aspects of "proximity" that can affect the outcome of collaborations. Physical proximity provides access to potential collaborators, and allows the collaboration seeker to make informal and unobtrusive assessments about compatibility. In the absence of physical proximity, shared research interests and/or common organizational or research communities can serve as surrogates.

(6) The system should facilitate the assessment of personal compatibility, similarity of work styles and other "soft" traits influencing collaborations.
Our background research indicated that personal compatibility and similar work style are important factors determining the success of collaborations. The literature also indicates that more than a simple overlap of interests is needed to create a successful collaboration. Expertise location systems should therefore facilitate an assessment of these factors, for instance, by identifying social connections.

(7) Social networks solely based on co-authorship may only partially describe a researcher's collaborative network.
Previous attempts to automatically describe a researcher's collaborative network based on co-authorship of papers were only partially successful [18,19]. Although co-authorship seems to be a good starting point for describing a collaboration network, it should be supplemented and validated by other data. Ideally, expertise location sys-

tems could create a preliminary network from co-authorship data that can be triangulated and validated using other information.

(8) The system should account for researchers' preferences regarding privacy and public availability of information about them.
To varying degrees, researchers tend to be protective of information about themselves or the projects they are working on. On the other hand, researchers are motivated to share information when they feel that doing so will add value to their work. As research on the structure and dynamics of networks has shown [20], central nodes in a network attract more links than peripheral nodes. By inference, highly productive scientists may be the focus of a disproportionately large number of contacts in professional networks. This type of social overload may cause them not to be favorable to additional contacts. Expertise location systems should therefore allow users to control whether they are visible at all, and, if so, which information is available about them under which circumstances.

(9) The system should provide methods to search effectively across disciplines.
Researchers need to be able to effectively search for collaborators in domains outside their own. However, researchers from one domain are unlikely to be aware of the terminology they need to search for in order to find a specific area of expertise. Standardized terminologies, such as Medical Subject Headings (MeSH), facilitate searching, but create artificial boundaries (for instance, between MeSH- and non-MeSH-indexed literatures). Systems that guide non-experts towards the appropriate subdomain and research category rather than making them provide keywords themselves may help ameliorate this problem.

(10) The system should help make "non-intuitive" connections between researchers.
Many scientific collaborations produce novel and innovative insights when the research interests of collaborators are complementary, or, at least, not closely aligned. However, similarity and complementarity of research interests are difficult to define. Multidisciplinary research is often viewed as requiring complementary expertise from different fields; however, even research teams within the same field are often configured to include investigators with slightly divergent interests. Many existing systems and resources focus on finding individuals with shared interests, which is much easier and straightforward than identifying those with complementary interests. One example for producing such connections computationally are systems that mine the literature for relations among research areas that are not obvious at first glance [21,22]. Advanced implementations of expertise location systems to support collaboration seekers could integrate such functionality.

3.2 The Semantic Web as a technical basis for expertise location systems

Few papers have discussed the problem of expertise location in the context of the Semantic Web [14,23-26]. However, Semantic Web technologies represent a rich array

of possibilities addressing many, but not all, of the requirements listed above. The Semantic Web is most likely to serve as a useful technological infrastructure for implementing expertise location systems, not as an end-to-end architecture.

Traditionally, a significant hurdle for adoption and use of expertise location systems has been the effort required to create and maintain comprehensive and up-to-date profiles. The Semantic Web can help ameliorate this problem by making information available that accumulates as a result of an individual's "digital activities." For instance, the Semantic Web makes it very easy to collate information from social networks and social media, for instance Friend-of-a-friend (FOAF) systems, online communities, blogs and information-sharing sites [14]. The resulting profile could, for instance, include topics that the individual has discussed with others or individuals s/he has interacted with. However, this information is not likely to substitute for more formal and rigorously maintained information, such as that found in a researcher's curriculum vitae (CV) [27]. Researchers in expertise location systems must clearly be motivated to keep their profile current, comprehensive and up-to-date, regardless of the method used to generate the data.

A related issue is the aggregation of data from sources other than the Web, for instance Collaborative Work Environments (CWEs). While CWEs tend to connect individuals within organizations quite well, they fail to do so among organizations. Information from CWEs made accessible through a framework such as Semantically Interlinked Online Communities (SIOC) [14] could contribute rich information to researcher profiles.

The Semantic Web also presents the opportunity to connect information created by an individual with information generated about the individual by others. MEDLINE, Google Scholar, the ISI Web of Science and other databases are examples of resources/databases that contain information about researchers. One significant challenge is to match information from different sources unambiguously to the individual. Ideally, the various online identities/unique identifiers of an individual are explicitly linked, as described by Bojars [14].

Automatically collating information using these strategies is likely to result in profiles that are more comprehensive and up-to-date than those compiled using other means. For instance, contributions to social media can be aggregated in near real-time and combined with information that may not be widely available in public for some time (for instance, a recently accepted paper listed in a CV). Social networks constructed from FOAF systems and online interactions may be more complete than or complementary to those based on co-authorship.

Expertise location systems need to be able to search across content domains as well as social spaces. Searching effectively across content domains requires ontologies, which are central to the vision of the Semantic Web [26]. While well-developed and sophisticated ontologies exist for some domains, for instance, the Medical Subject Headings used to index the biomedical literature, they are not universally available. Semantic mapping among different ontologies is a significant problem on which considerable attention has been focused [28-30]. Computational tools to bridge queries among different ontologies have been described [23,28], but at present, no large-scale trials examining how well the approach works in practice (similar, for instance, to the National Library of Medicine's Large Scale Vocabulary Test [31]) have been published.

Searching across social spaces suffers from a similar problem if individual identities can not be matched between systems. Frameworks such as OpenSocial [32] are essential to allowing users to traverse social networks without regard to system boundaries.

Building expertise location systems on top of the Semantic Web does not only require the capability to aggregate and organize data about each expert, but also to present the data in a usable and useful form to collaboration seekers. The experts listed by the system must be able to view and, if necessary, change how they appear to users of the system. This includes taking individual needs for controlled access to profile information into account. For instance, researchers may prefer to limit public, anonymous access to information about them, but be more open within their social network. Second, systems should facilitate rapid, progressively detailed review of potential collaborators. Given the fact that choosing a collaborator is a highly subjective and idiosyncratic process, system performance may be weighted to provide a larger number of potential candidates, rather than attempting to present only a few candidates that the system perceives as "optimal." This tradeoff between sensitivity and specificity could be adjusted as the system learns about the preferences of its users.

In summary, the Semantic Web presents many opportunities for helping implement expertise location systems. However, the Semantic Web does not exist independent of the computational tools, environment, workflow and user behavior of biomedical scientists, and thus must integrate with the current context of system use, not strive to replace it.

4 Discussion

Given the increasing trend towards collaboration in science, as well as the expanding universe of potential collaborators for scientists, electronic systems can be expected to play an increasingly important role in connecting scientists to one another. While traditional approaches will always play a role in how scientists connect with and select collaborators, expertise location systems have the potential to improve how effectively and efficiently scientists form collaborations. At their lowest level of implementation, they can reduce the workload of simple tasks related to forming collaborations, for instance collecting and organizing information about a potential collaborator. More advanced functionality would allow collaboration seekers to use information not usually available to them, for instance how potential collaborators relate to the seeker's existing social network. Further developments could integrate computational approaches to identifying scientific opportunities, as Swanson has demonstrated [22].

Our research has shown that expertise location systems for establishing collaborations in biomedical science have a complex and multifaceted set of requirements. Clearly, one challenge for designing these systems is that seeking, evaluating and choosing scientific collaborators is a complex decision-making process that is poorly understood. Our study only presents a first step in understanding how to build systems that are truly useful tools for establishing promising and high-impact collaborations. The list of requirements we formulated is clearly preliminary, and should be validated

with a larger number of participants, at other institutions/research settings and in other geographic locations. A competitive analysis of existing systems may have provided additional and useful formal data to this study. However, the rapidly moving market for such systems would have reduced the usefulness of such an evaluation beyond a very limited time frame.

A related question is how well the requirements, which are mainly based on findings from biomedical disciplines, generalize to other scientific domains. While we drew on literature that included studies from a variety of scientific disciplines, our observations and interviews were conducted predominantly with biomedical scientists. Therefore, claims of generalizability are difficult to make, especially given the specific history, culture and structure of the biomedical research enterprise in the US. For instance, federal funding agencies, such as the NIH, play a very prominent role in shaping researcher behavior and priorities. (The current trend towards multidisciplinary research is an example.) Second, non-research oriented organizations, such as for-profit hospital systems, function both as data providers and employers of some researchers. This circumstance can influence collaborative behavior, for instance when the organization attempts to preserve its competitive advantage through policies limiting collaboration. Clearly, the history and tradition of collaborative work in a discipline can influence individual behavior. As a recent book suggests [33], some research areas, such as high-energy physics and astronomy, have a much stronger tradition of collaboration and data sharing than other fields. While the requirements articulated in this paper may be seen as a viable starting point, additional work is needed to understand the degree to which they can be generalized.

Additional studies, both in biomedical science and in other fields, should also be helpful in elucidating some of the implicit contradictions in the current list of requirements. For instance, the desire for privacy of selected information (Requirement 8) conflicts, to some degree, with the need to provide comprehensive information (Requirement 2) and the desire to search effectively across disciplines (Requirement 9). The trade-offs among the requirements are likely context-dependent, and further research should provide insight into situations and use cases where and how particular trade-offs should be made.

As shown above, Semantic Web technologies have significant potential for addressing the requirements for expertise location systems. Integrating information from disparate and inhomogeneous sources using ontologies and annotation frameworks are key to creating the rich and comprehensive profiles that are the basis for making connections among researchers. Several challenges present themselves for future research in this context. First, we need to understand in more depth how scientists seek, evaluate and choose evaluators. Such research should include, for instance, factors that motivate and prompt scientists to look for collaborators; the criteria they use to evaluate them; and circumstances influencing the adoption of new tools to support the formation of collaboration. Second, we need to begin the process of translating system requirements into Semantic Web applications. Early efforts in this area have been encouraging [13,23]. However, we need to ensure that these applications work in a generalizable manner, and do not result in insular applications that are difficult to apply in other contexts. Third, we need to begin to consider measurements for system performance of expertise location systems. Analogously to benchmarking systems in information retrieval, we need to define performance criteria and system outcomes.

What constitutes a "relevant hit" in an expertise location system? How does relevance vary based on different user characteristics? What role do semantic technologies play in achieving and assessing system outcomes? As we address these research questions, expertise location systems have the potential to become increasingly important in enhancing and strengthening scientific collaboration.

Acknowledgments
This project was, in part, supported by grant UL1 RR024153 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research. We appreciate Ellen Detlefsen's and Erin Nordenberg's help with interviewing scientists, Janice Stankowicz's help with managing the interview process, and Michael Dziabiak's help with formatting and submission. Special thanks go to the reviewers and their constructive comments, which helped improve the paper significantly.

Daniel Weiss's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.

References

1. Braun, T., Schubert, A.: A Quantitative View on the Coming of Age of Interdisciplinarity in the Sciences 1980-1999. Scientometrics. 58, 183-189 (2003)
2. Moses, H. III., Dorsey, E.R., Matheson, D.H., Their, S.O.: Financial Anatomy of Biomedical Research. JAMA. 294, 1333-1342 (2005)
3. O'Dell, C., Grayson Jr., C.J.: If Only We Knew What We Know: The Transfer of Internal Knowledge and Best Practice. Free Press, New York (1998)
4. Stenmark, D.: Leveraging Tacit Organizational Knowledge. J. Manage. Inform. Syst. 17, 9-24 (2000)
5. Kraut, R.E., Galegher, J., Egido, C.: Relationships and Tasks in Scientific Research Collaboration. Hum-Comput. Interact. 3, 31-58 (1987-1988)
6. McDonald, D.W., Ackerman, M.S.: Expertise Recommender: A Flexible Recommendation System and Architecture. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 231-240. ACM, New York (2000)
7. Erickson, T., Kellogg, W.A.: Knowledge Communities: Online Environments for Supporting Knowledge Management and Its Social Context. In: Ackerman, M.S., Pipek, V., Wulf, V. (eds.) Sharing Expertise: Beyond Knowledge Management. pp. 299-325. MIT Press, Cambridge (2003)
8. Millen, D.R., Fontaine, M.A., Muller, M.J.: Understanding the Benefit and Costs of Communities of Practice. Commun. ACM. 45, 69-73 (2002)
9. Ackerman, M.S., Palen, L.: The Zephyr Help Instance: Promoting ongoing Activity in a CSCW System. In: Tauber, M.J. (ed.) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground, pp. 268-275. ACM, New York (1996)
10. Jacovi, M., Soroka, V., Ur, S.: Why Do We ReachOut?: Functions of a Semi-Persistent Peer Support Tool. In: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, pp. 161-169. ACM, New York (2003)
11. Friedman, P.W., Winnick, B.L., Friedman, C.P., Mickelson, P.C.: Development of a MeSH-based Index of Faculty Research Interests. Proc. AMIA. Symp., 265-269 (2000)

12. Mockus, A., Herbsleb, J.D.: Expertise Browser: A Quantitative Approach to Identifying Expertise. In: Proceedings of the 24th International Conference on Software Engineering, pp. 503-512. ACM, New York (2002)
13. Sriharee, N., Punnarut, R.: Constructing Semantic Campus for Academic Collaboration. In: Zhdanova, A.V., Nixon, L.J.B., Mochol, M., Breslin, J.G. (eds.) Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics, pp. 23-32. (2007)
14. Bojars, U., Breslin, J.G., Peristeras, V., Tummarello, G., Decker, S.: Interlinking the Social Web with Semantics. IEEE. Intell. Syst. 23, 29-40 (2008)
15. Beyer, H., Holtzblatt, K.: Contextual Design: Defining Customer-Centered Systems. Morgan Kaufmann, San Francisco (1998)
16. Glaser, B.G., Strauss, A.L.: The Discovery of Grounded Theory; Strategies for Qualitative Research. Aldine Pub. Co., Chicago (1967)
17. Kautz, H., Selman, B., Shah, M.: The Hidden Web. AI. Mag. 18, 27-26 (1997)
18. Bordons, M., Gómez, I.: Collaboration Networks in Science. In: Cronin, B., Atkins, H.B. (eds.). The Web of Knowledge: A Festschrift in Honor of Eugene Garfield. pp. 197-213. Information Today, Medford (2000)
19. Katz, J.S., Martin, B.R.: What Is Research Collaboration? Res. Policy. 26, 1-18 (1997)
20. Newman, M., Barabási, A.L., Watts, D.J., editors.: The Structure and Dynamics of Networks. Princeton University Press, Princeton (2006)
21. Bahr, N., Cohen, A.M.: Discovering Synergistic Qualities of Published Authors to Enhance Translational Research. Proc. AMIA. Symp., in press (2008)
22. Swanson, D.R., Smalheiser, N.R.: An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery. Artif. Intell. 91, 183-203 (1997)
23. Liu, P., Dew, P.: Using Semantic Web Technologies to Improve Expertise Matching within Academia. In: I-KNOW '04, pp. 370-378. (2004)
24. Liu, P., Curson, J., Dew, P.: Use of RDF for Expertise Matching within Academia. Knowl. Inf. Syst. 8, 103-30 (2005)
25. Cameron, D., Aleman-Meza, B., Arpinar, I.B.: Collecting Expertise of Researchers for Finding Relevant Experts in a Peer-Review Setting. In: 1st International ExpertFinder Workshop, (2007)
26. Davies, J., Duke, A., Sure, Y.: OntoShare - An Ontology-Based Knowledge Sharing System for Virtual Communities of Practice. J. Univers. Comput. Sci. 10, 262-283 (2004)
27. Schleyer, T., Spallek, H., Butler, B.S., Subramanian, S., Weiss, D., Poythress, M.L., et al.: Facebook for Scientists: Requirements and Services for Optimizing How Scientific Collaborations Are Established. J. Med. Internet. Res. Forthcoming (2008)
28. Doan, A., Madhavan, J., Dhamankar, R., Domingos, P., Halevy, A.: Learning to Match Ontologies on the Semantic Web. VLDB J 12, 303-319 (2003)
29. Castano. S., Ferrara, A., Montanelli, S.: Matching Ontologies in Open Networked Systems: Techniques and Applications. Journal on Data Semantics V, 25-63 (2006)
30. Lee, Y., Supekar, K., Geller, J.: Ontology Integration: Experience with Medical Terminologies. Comput. Biol. Med. 36, 893-919 (2006)
31. Humphreys, B.L., McCray, A.T., Cheh, M.L.: Evaluating the Coverage of Controlled Health Data Terminologies: Report on the Results of the NLM/AHCPR Large Scale Vocabulary Test. J. Am. Med. Inform. Assoc. 4, 484-500 (1997)
32. Mitchell-Wong, J., Kowalczyk, R., Roshelova, A., Joy, B., Tsai, H.: OpenSocial: From Social Networks to Social Ecosystem. In: Proceedings of the 2007 Inaugural IEEE International Conference on Digital Ecosystems and Technologies, pp. 361-366. (2007)
33. Olson, G.M., Zimmerman, A., Bos, N.: Scientific Collaboration on the Internet. MIT Press, Cambridge (2008)

Smushing RDF instances: are Alice and Bob the same open source developer?

Lian Shi[1], Diego Berrueta[1], Sergio Fernández[1],
Luis Polo[1], and Silvino Fernández[2]

[1] Fundación CTIC
Gijón, Asturias, Spain
{firstname.lastname}@fundacionctic.org
http://www.fundacionctic.org/
[2] R&D Technological Centre (CDT)
ArcelorMittal Asturias
Avilés, Asturias, Spain
silvino.fernandez@arcelormittal.com
http://www.arcelormittal.com/

Abstract. Analysing RDF data gathered from heterogeneous Semantic Web sources requires a previous step of consolidation in order to remove redundant instances (data smushing). Our aim is to explore and integrate smushing techniques to improve recall, i.e., to find as many redundant instances as possible. Two approaches to spot resources with the same identity are described: the first one is based on Logics, exploiting OWL inverse functional properties (IFP); the second one is based on traditional IR techniques, e.g., resource label comparison. We evaluate experimental results in the context of open source communities.

1 Introduction

The increasing amount of machine processable data in the Semantic Web facilitates processes such as social network analysis and data mining. Innovative applications, like expert finding on the (Semantic) Web, are enabled by the ability of executing these processes at a World-Wide Web scale. Although the RDF data model is well suited to seamlessly merge data (triples) from arbitrary sources, a data integration problem still remains. Unconnected descriptions of the same thing can be obtained from different sources. For instance, a single individual can participate in several web communities with different virtual identities. When they are summed together, the descriptions of her virtual identities (such as e-mail accounts) will be different RDF resources weakly connected to each other. If these identities were to be taken as different persons, data analysis would be crippled, as it would lead to imprecise conclusions and a widespread flooding of phantom virtual identities.

Social communities, their networks and their collaborative forums, are one particular focus of interest for analysis, as they provide large amounts of data

that can be used for several purposes. People in these communities share common interests, exchange information and interact with each other. FOAF [3] (short for Friend-Of-A-Friend) and SIOC [2] (Semantically-Interlinked Online Communities) offer vocabularies for publishing machine readable descriptions of people, making it possible to link from one site, person, company, etc. to related ones. With their popularity and wide acceptance as a de facto standard vocabularies for representing social networks, there is a dynamic increase in the amount of social profiles available in these formats produced by many large social networking websites. This fact can be verified by the number of documents that use these namespaces in the Semantic Web [6].

We use data mined from open source communities to evaluate two smushing techniques in order to merge the virtual identities of the members of these communities. That is, we aim to identify the co-occurrence of the same person in different communities[3]. The first approach exploits the semantics of inverse functional properties, which solely and definitely determines whether two entities are the same considering their property values. The second approach is not based on Logics, but on heuristics, more precisely, on the comparison of entity labels. Both techniques are applied to a dataset that contains thousands of instances of `foaf:Person`.

The paper is structured as follows: next, we briefly introduce the most important related work. We detail two complementary smushing strategies in Section 3. Section 4 describes how a corpus of RDF data was collected from some communities that focus on open source software development. Experimental results are exposed in Section 5, and their interpretation in the context of open source communities is discussed in Section 6. Finally, Section 7 closes the paper with conclusions on the smushing process and some insights into future work.

2 Related Work

Social networks have opened up a new sight because people provide information about themselves and their social connections in publicly accessible forums. The main topics and subjects of a vast literature of previous works about social networks include examinations of online social networks such as [10], which recommends a survey-based approach for extracting social information about users. Their growth and activity patterns, design and behaviour in online communities has also been studied [11, 20].

In [14], the authors show how large isolated data graphs from disparate structured data sources can be combined to form one, large, well-lined RDF graph. Their work provides a large corpus that can act as a benchmark dataset for evaluating expert finding algorithms, and it also simulates the availability of real-world data used in various research scenarios.

Ding et al. [5] present a novel perspective of the Semantic Web of FOAF documents, and proposed a heuristic approach to identify and discover FOAF

[3] In this sense our research relates to matching frameworks, see `http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining`

documents from the Web and to extract information about people from these documents. Their work can be used to discover existing and emerging online communities.

The application of machine learning technologies to FOAF has also been explored, highlighting the challenges posed by the characteristics of such data. The authors of [12] experiment with profiles and generate a set of rules for adding properties to users found to be in a set of clusters, and also for learning descriptions of these groups. They argue these descriptions can be used later for on-the-fly personalisation tasks.

3 RDF data smushing

We call *smushing* to the process of normalising an RDF dataset in order to unify *a priori* different RDF resources which actually represent the same thing[4]. The application which executes a *data smushing* process is called a *smusher*. The process comprises two stages: first, redundant resources are identified; then, the dataset is updated to reflect the recently acquired knowledge. The latter is usually achieved by adding new triples to the model to relate the pairs of redundant resources. The `owl:sameAs` is often used for this purpose, although other properties without built-in logic interpretations can be used as well (e.g.: `ex:hasSimilarName`). We will expand on this at the end of this section.

Redundant resources can be spotted using a number of techniques. In this paper, we explore two of them: (1) using logic inference and (2) comparing labels. We note that other approaches are possible as well, including custom rule-based systems, human computation and user-contributed interlinking (UCI) [13].

3.1 Inverse Functional Properties

OWL [18] introduces a kind of object properties called *Inverse Functional Properties* (`owl:InverseFunctionalProperty`, IFPs for short). An IFP is a property which behaves as an injective association, hence its values uniquely identify the subject instance:

$$\forall p/p \in \text{IFP} \Rightarrow (\forall s_1, s_2 \,/\, p(s_1) = p(s_2) \Rightarrow s_1 = s_2) \tag{1}$$

This inference rule is built-in in the OWL-DL reasoners, therefore, this kind of smushing can be easily achieved just by reasoning the model. However, it is advisable to avoid the reasoner and to implement the IFP semantics by means of an *ad hoc* rule. These are the reasons:

– Executing a simple, light-weight rule is often more efficient than the reasoner, which usually performs many other tasks. Moreover, it can be used regardless of the expressivity level of the dataset, while reasoners have unpredictable behaviour for OWL-Full datasets.

[4] This use of the expression *data smushing* in this context can be traced back to Dan Brickley: `http://lists.w3.org/Archives/Public/www-rdf-interest/2000Dec/0191.html`

```
CONSTRUCT {
  ?person1 owl:sameAs ?person2
}
WHERE {
  ?person1 rdf:type  foaf:Person .
  ?person2 rdf:type  foaf:Person .
  ?person1 foaf:mbox_sha1sum ?email .
  ?person2 foaf:mbox_sha1sum ?email .
  FILTER (?person1 != ?person2)
}
```

Fig. 1: IFP smushing rule implemented as a SPARQL CONSTRUCT sentence. The usual namespace prefixes are assumed.

– A custom rule can generalise IFPs to any kind of properties, including datatype properties. There are some scenarios in which such generalisation is useful. For instance, while the object property `foaf:mbox` is declared as an IFP, whose value is often unavailable due to privacy concerns. On the other hand, values of the property `foaf:mbox_sha1sum` are widely available (or can be easily calculated from the former), but as it is a datatype property, it cannot be declared an IFP in OWL.

This rule can be written as a SPARQL CONSTRUCT sentence, according to the idiom described by [19], see Figure 1. Note that this rule only takes into account the `foaf:mbox_sha1sum` property, but its generalisation to any property declared as `owl:InverseFunctionalProperty` is straightforward.

When smushing resources that describe people, some FOAF properties can be used as IFPs. The FOAF specification defines `mbox`, `jabberID`, `mbox_sha1sum`, `homepage`, `weblog`, `openid` as IFP, among others. However, a quick analysis of a set of FOAF files collected from the web shows that some of these properties are barely used, while others are often (mis-)used in a way that makes them useless as IFP. Notably, some users point their `homepage` to their company/university homepage, and `weblog` to a collective blog. Therefore, we restrict our smusher to the `mbox_sha1sum` property.

3.2 Label similarity

The concept of similarity is extensively studied in Computer Science, Psychology, Artificial Intelligence, and Linguistics literature. String similarity plays a major role in Information Retrieval. When smushing people's descriptions, labels are personal names (`foaf:name`). Nevertheless, personal names follow particular rules which make them intractable. Personal names, as any other word, can be miss-spelled, however, the probability of such error is very low if the names are entered by their owners. Therefore, traditional similarity comparison functions,

such as Levenshtein distance, are not really useful. In [4] the authors describe advanced techniques for personal name comparison, and in [17] study their impact on the precision. We just use a much simpler strict string equality comparision, ignoring common invalid names often found in email headers.

Smushing based on label similarity deals with imprecise knowledge, i.e., even a perfect label equality does not guarantee that two resources are the same. Using a softer comparison function will produce even more uncertain knowledge.

A label-based smusher can be implemented as a rule. Unfortunately, SPARQL does not have rich built-in string comparison functions. There is a proposed extension call iSPARQL [16] that can be used with this purpose. Our experience reveals that iSPARQL implementation is far from being efficient enough to deal with large datasets. This fact suggests that other approaches to implement label-based smushing should be considered.

3.3 Smushing, correctness and consistence

The pairs of redundant resources identified using the techniques described above can be used to enrich the dataset. OWL provides a special property to "merge" identical resources, `owl:sameAs`. When two resources are related by `owl:sameAs`, they effectively behave as a single resource for all the OWL-aware applications. Note, however, that plain SPARQL queries operate at the RDF level, and therefore they are unaware of the `owl:sameAs` semantics.

Anyway, the semantics of `owl:sameAs` may be too strong for some cases. On the one hand, some applications may still want to access the resources individually. On the other hand, several factors can influence on the reliability of the findings made by the smusher. Notably, the data smushing based on label comparison is obviously imperfect, and can lead to incorrect results. For instance, different people can have the same name, or they can fake their identities. Even the logically-sound smushing based on IFPs is prone to error, due to the low-quality of the input data (fake e-mail addresses, identity theft). Although improbable, it is also possible that different e-mail addresses clash when they are hashed using SHA1 [7].

To tackle these issues, a custom property can be used instead, such as `ex:similarNameTo`. Applications interested in the strong semantics of `owl:sameAs` can still use a rule to re-create the links.

Another kind of OWL properties, Functional Properties (FP) are also useful for smushing. They can help to check the consistency of the smusher's conclusions. A resource cannot have multiple different values for a FP. Therefore, if two resources that are to be smushed are found to have irreconcilable values for `foaf:birthday`, an issue with the smushing rules (or the quality of the input data) must be flagged.

4 Data recollection

A corpus of RDF data with many `foaf:Person` instances was assembled by crawling and scrapping five online communities. There is a shared topic in these

communities, namely open source development, hence we expect them to have a significative number of people in common.

We continue the work started in [1] to mine online discussion communities, and we extend it to new information sources. We use the following sources:

- GNOME Desktop mailings lists: all the authors of messages in four mailing lists (*evolution-hackers*, *gnome-accessibility-devel*, *gtk-devel* and *xml*) within the date range July 1998 to June 2008 were exported to RDF using SWAML [9].
- Debian mailing lists: all the authors of messages in four mailing lists (*debian-devel*, *debian-gtk-gnome*, *debian-java* and *debian-user*) during years 2005 and 2006 were scrapped from the HTML versions of the archives with a set of XSLT style sheets to produce RDF triples.
- Advogato: this community exports its data as FOAF files. We used an RDF crawler starting at Miguel de Icaza's profile. Although Advogato claims to have +13,000 registered users, only +4,000 were found by the crawler.
- Ohloh: the RDFohloh project [8] exposes the information from this directory of open source projects and developers as Linked Data. Due to API usage restrictions, we could only get data about the +12,000 oldest user accounts.
- Debian packages: descriptions of Debian packages maintainers were extracted from *APT* database of Debian packages in the *main* section of the *unstable* distribution[5].

Instances generated from these data sources were assigned a URI in a different namespace for each source. Some of these data sources do not directly produce instances of `foaf:Person`, but just instances of `sioc:User`.An assumption is made that there is a `foaf:Person` instance for each `sioc:User`, with the same e-mail address and name. These instances were automatically created when missing. This assumption obviously leads to redundant instances of `foaf:Person` which will be later detected by the smusher.

5 Experimental Results

The ultimate goal of our experiments is to exercise the smushing processes described in Section 3 against a realistic dataset. Two million RDF triples were extracted from the sources described above, and put into OpenLink Virtuoso server[6] which provides not only an effective triple store, but also a SPARQL end-point that was used to execute queries using scripts. Table 1a summarises the number of instances of `foaf:Person` initially obtained from each source.

We evaluated two smushers: the first one smushed `foaf:Person` instances assuming that `foaf:mbox_sha1sum` is an IFP; the second one smushed the same instances comparing their `foaf:name` labels for string strict equality, without any

[5] Retrieved August 3rd, 2008.
[6] `http://virtuoso.openlinksw.com/`

Source	`foaf:Person` instances
DebianPkgs	1,845
Advogato	4,168
GnomeML	5,797
Ohloh	12,613
DebianML	12,705

(a)

	Num. of people
In 5 communities	1
In 4 communities	37
In 3 communities	273
In 2 communities	1,669

(b)

Table 1: Size of the studied communities before smushing (1a), and people accounting by the number of communities they are present in, after smushing (1b).

normalisation. Both smushers were implemented using SPARQL CONSTRUCT rules. The newly created `owl:sameAs` triples were put in different named graphs.

These links were analysed to find co-occurrences of people in different communities. The absolute co-occurrence figures are presented in Tables 2 and 3, respectively. Later, the two named graphs were aggregated. Table 4 contains the absolute co-occurrence considering the combined results of both smushers. Note that some redundancies are detected by both smushers, therefore figures in Table 4 are not the sum of two previous ones. These matrices are symmetrical, hence we skip the lower triangle.

The degree of overlap between communities is better observed in Tables 5, 6 and 7, which present the ratio of overlap relative to the size of each community.

Tables 1b and 8 study the number of communities each person is present in. Interestingly enough, an individual was found to have presence in all five communities.

6 Discussion

The elements of the main diagonal of Tables 2, 3, 4 show the overlap within each community, i.e., the number of people that have registered more than once in each community. Some communities use the e-mail address as the primary key to identify their users, therefore, the smushing process using the e-mail as IFP (Table 2) has zeros in the main diagonal for these communities. However, other communities use a different primary key, thus allowing users to repeat their e-mail addresses. For instnace, a small number of users have registered more than one account in Advogato with the same e-mail (these account have been manually reviewed, and they seem to be accounts created for testing purposes).

Our data acquisition process introduces a key difference between how user accounts are interpreted in Debian mailing lists and GNOME mailing lists. The former considers e-mail address as globally unique, i.e., the same e-mail address posting in different Debian mailing lists is assumed to belong to the same user. On the other hand, a more strict interpretation of how Mailman works is made with respect to the GNOME mailing lists, where identical e-mail address posting in different mailing lists are assumed to belong to a priori different users. In the

Source	DebianPkgs	Advogato	GnomeML	Ohloh	DebianML
DebianPkgs	0	81	37	74	762
Advogato		19	270	106	141
GnomeML			364	112	161
Ohloh				0	115
DebianML					0

Table 2: Number of smushed instances of `foaf:Person` using `foaf:mbox_sha1sum` as IFP.

Source	DebianPkgs	Advogato	GnomeML	Ohloh	DebianML
DebianPkgs	98	170	101	58	1319
Advogato		49	592	95	305
GnomeML			1716	148	432
Ohloh				13	208
DebianML					2909

Table 3: Number of smushed instances of `foaf:Person` with exactly the same `foaf:name`.

Source	DebianPkgs	Advogato	GnomeML	Ohloh	DebianML
DebianPkgs	98	188	113	104	1418
Advogato		55	669	167	342
GnomeML			1765	227	462
Ohloh				13	287
DebianML					2909

Table 4: Number of smushed instances of `foaf:Person` combining the two smushing techniques.

Source	DebianPkgs	Advogato	GnomeML	Ohloh	DebianML
DebianPkgs	0.00%	4.39%	2.01%	4.01%	41.30%
Advogato	1.94%	0.46%	6.48%	2.54%	3.38%
GnomeML	0.64%	4.66%	6.28%	1.93%	2.78%
Ohloh	0.59%	0.84%	0.89%	0.00%	0.91%
DebianML	6.00%	1.11%	1.27%	0.91%	0.00%

Table 5: Ratio of smushed instances of `foaf:Person` using IFP, relative to the size of the community of the row.

Source	DebianPkgs	Advogato	GnomeML	Ohloh	DebianML
DebianPkgs	5.31%	9.21%	5.47%	3.14%	71.49%
Advogato	4.08%	1.18%	14.20%	2.28%	7.32%
GnomeML	1.74%	10.21%	29.60%	2.55%	7.45%
Ohloh	0.46%	0.75%	1.17%	0.10%	1.65%
DebianML	10.38%	2.40%	3.40%	1.64%	22.90%

Table 6: Ratio of smushed instances of `foaf:Person` with exactly the same `foaf:name`, relative to the size of the community of the row.

Source	DebianPkgs	Advogato	GnomeML	Ohloh	DebianML
DebianPkgs	5.31%	10.19%	6.12%	5.64%	76.86%
Advogato	4.51%	1.32%	16.05%	4.01%	8.21%
GnomeML	1.95%	11.54%	30.45%	3.92%	7.97%
Ohloh	0.82%	1.32%	1.80%	0.10%	2.28%
DebianML	11.16%	2.69%	3.64%	2.26%	22.90%

Table 7: Ratio of smushed instances of `foaf:Person` combining both techniques, relative to the size of the community of the row.

Name	Number of Name vars.	Number of E-mail acc.	Presence in DebianPkgs	Advogato	GnomeML	Ohloh	DebianML
Frederic P.	1	3	•	•	•	•	•
Dan K.	2	3	○	•	•	•	•
Jerome W.	2	1	•	○	•	•	•
Raphael H.	2	3	•	○	•	•	•
Person #01	1	1	•	○	•	•	•
Person #02	1	1	•	○	•	•	•
Person #03	2	4	•	○	•	•	•
Julien D.	1	4	•	•	○	•	•
Rob B.	1	2	•	•	○	•	•
Daniel R.	2	1	•	•	○	•	•
Gürkan S.	5	2	•	•	○	•	•
Ricardo M.	2	3	•	•	○	•	•
Ray D.	3	2	•	•	○	•	•
Person #04	1	3	•	•	○	•	•
Person #05	1	3	•	•	○	•	•
Person #06	2	5	•	•	○	•	•
Person #07	1	3	•	•	○	•	•
Person #08	2	2	•	•	○	•	•
Person #09	1	2	•	•	○	•	•
Person #10	2	3	•	•	○	•	•
Federico Di G.	1	2	•	•	•	○	•
Ross B.	1	2	•	•	•	○	•
Person #11	1	5	•	•	•	○	•
Person #12	1	2	•	•	•	○	•
Person #13	1	3	•	•	•	○	•
Person #14	1	2	•	•	•	○	•
Person #15	1	2	•	•	•	○	•
Person #16	1	2	•	•	•	○	•
Person #17	1	2	•	•	•	○	•
Person #18	1	1	•	•	•	○	•
Person #19	1	3	•	•	•	○	•
Francis T.	2	3	•	•	•	•	○

Table 8: Details of the top people by the number of communities they are present in. A filled dot (•) denotes presence in the community. In order to protect privacy, we only print real names of people who have given us explicit permission to do so.

second case, we rely on the smushing process to merge the identities of these users. The number of smushed instances in Table 2 evidence the fact that people post messages to different mailing lists using the same e-mail address.

Although they must be handled with extreme care due to the issues aforementioned, the combined results of the two smushing processes are consistent with the expected ones. For instance, there is a very high overlap between the Debian developers (maintainers of Debian packages) and the Debian mailing lists. Obviously, Debian developers are a relatively small group at the core of the Debian community, thus they are very active in its mailing lists. Another example is the overlap between Advogato and GNOME mailing lists. Advogato is a reputation-based social web site that blossomed at the same time that the GNOME project was gaining momentum. Advogato was passionately embraced by the GNOME developers, who used Advogato to rate each others' development abilities.

We also studied whether there are some people that are present in many of the communities at the same time. We chose communities which are closely related to each other, consequently, we expected a high number of cross-community subscribers. Table 1b evidences that there are several people who are present in many communities. From Table 8 we conclude that almost all the most active open source developers in our dataset are core members of the Debian community. Another interesting fact is that only a few people among the top members of the communities consistently use a single e-mail address and just one variant of their names. This fact proves the difficulty of the smushing process, but also its usefulness.

7 Conclusions and Future Work

In this paper, we explored smushing techniques to spot redundant RDF instances in large datasets. We have tested these techniques with more than 36,000 instances of `foaf:Person` in a dataset automatically extracted from different online open source communities. We have used only public data sources, consequently, these instances lack detailed personal information.

Comparing the figures in Tables 5 and 6, it is clear that the label-based smusher draws more conclusions than the IFP-based one. The number of redundant resources detected by the former is almost always higher than the one detected by the latter. Moreover, when compared to the aggregated figures in Table 7, we observe that the conclusions of the IFP-based smusher are largely contained in the conclusions of the label-based one. This fact can be explained because users with the same e-mail address often happen to have the same name. The difference between figures in Table 6 and 5 are explained by two facts: (a) there are people who have more than one e-mail account, and (b) there are different people with the same name (namesake). Unfortunately, it is not clear which is the influence of each factor. This is an issue, as smushing conclusions derived from (b) are obviously incorrect.

We are aware of the extreme simplicity of our experimentation using label comparison. In our opinion, however, it contributes to show the potential of this smushing technique. We note that it is possible to have more usages for it, for instance, smushing not just by people's names, but also by their publications, their organisations, etc. Surprisingly, the named-based smushing finds a high number of redundant resources even if the comparison strategy for labels (names) is very simplistic (in this case, case-sensitive string equality comparison). More intelligent comparison functions should lead to a higher recall. In this direction, we are evaluating some normalisation functions for names. We have also evaluated classical IR comparison functions that take into account the similarity of the strings (e.g., Levenshtein); nevertheless, their applicability to compare people's names is open to discussion. In general, a smusher algorithm has a natural maximum complexity of $O(n^2)$ due to the need to compare every possible pair of resources. This complexity raises some doubts about their applicability for very large dataset. Generalisation of these techniques to a web-scale will require to find ways to cut down the complexity.

We believe that the ratio of smushing can be further improved if the dataset is enriched with more detailed descriptions about people. Experiments are being carried out to retrieve additional RDF data from semantic web search engines such as SWSE [15] and Sindice [21] as a previous step to smushing. However, this work is still ongoing and we expect to present it in an upcoming publication. We aim to repeat our experiments in other communities apart from the open source one, for instance the Semantic Web community. The ExpertFinder Corpus [14] and the Semantic Web Conference Corpus[7] can be used for this purpose.

We intend to use our smusher to further investigate the potential for optimisations of the smushing process. The way in which these techniques are implemented is critical to achieve a promising performance of the smushing process, specially for very large datasets. In parallel, increasing the precision of smushing will require to study how to enable different smushing strategies to interrelate and reciprocally collaborate. We have started contacts with people from Table 8 asking them to confirm the communities they participate in; we will use their feedback to to measure the recall and the precision of the smushing process.

We acknowledge that any work on data mining, and in particular, identity smushing, raises some important privacy issues and ethical questions, even when the data used is publicly available on the Web. Actually we got very negative feedback from one the top members of open source communities, tagging this research topic as "*immoral*". Obviously we do not share this point of view, but we understand the privacy issues behind this opinion and we have tried to be extremely careful with the personal information that we manage and print.

References

1. D. Berrueta, S. Fernández, and L. Shi. Bootstrapping the Semantic Web of Social Online Communities. In *Proceedings of workshop on Social Web Search and Mining*

[7] `http://data.semanticweb.org/`

(SWSM2008), co-located with WWW2008, Beijing, China, April 2008.
2. J. Breslin, S. Decker, A. Harth, and U. Bojars. SIOC: an approach to connect web-based communities. *International Journal of Web Based Communities*, 2006.
3. D. Brickley and L. Miller. FOAF: Friend-of-a-Friend. `http://xmlns.com/foaf/0.1/`, November 2007.
4. W. W. Cohen, P. Ravikumar, and S. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of the workshop on Information Integration on the Web*, pages 73–78, 2003.
5. L. Ding, T. Finin, and A. Joshi. Analyzing social networks on the semantic web. *IEEE Intelligent Systems*, 9, 2005.
6. L. Ding, L. Zhou, T. Finin, and A. Joshi. How the semantic web is being used: An analysis of FOAF documents. In *38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, Track 4. IEEE Computer Society, 2005.
7. D. Eastlake and P. Jones. RFC 3174: US Secure Hash Algorithm 1 (SHA1). Technical report, IETF, 2001.
8. S. Fernández. RDFohloh, a RDF wrapper of Ohloh. In *1st workshop on Social Data on the Web (SDoW2008), co-located with ISWC2008*, Karlsruhe, Germany, October 2008.
9. S. Fernández, D. Berrueta, and J. E. Labra. Mailing lists meet the Semantic Web. In *Proceedings of 1st workshop on Social Aspects of the Web (SAW2007), co-located with BIS2007*, Poznan, Poland, April 2007.
10. L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. *Journal of Computer Mediated Communication*, 3, 1997.
11. J. Golbeck and M. Rothstein. Linking Social Networks on the Web with FOAF. In *Proceedings of the 17th international conference on World Wide Web (WWW2008)*, 2008.
12. G. A. Grimnes, P. Edwards, and A. Preece. Learning meta-descriptions of the FOAF network. In *3rd International Semantic Web Conference (ISWC2004)*, 2004.
13. M. Hausenblas, W. Halb, and Y. Raimond. Scripting User Contributed Interlinking. In *Proceedings of the 4th workshop on Scripting for the Semantic Web (SFSW2008), co-located with ESWC2008*, Tenerife, Spain, June 2008.
14. A. Hogan and A. Harth. The ExpertFinder Corpus 2007 for the Benchmarking and Development of Expert-Finding Systems. In *Proceedings of 1st International ExpertFinder Workshop*, 2007.
15. A. Hogan, A. Harth, J. Umrich, and S. Decker. Towards a Scalable Search and Query Engine for the Web. In *Proceedings of the 16th international conference on World Wide Web (WWW2007)*, pages 1301–1302, New York, NY, USA, 2007.
16. C. Kiefer. Imprecise SPARQL: Towards a Unified Framework for Similarity-Based Semantic Web Tasks. In *2nd Knowledge Web PhD Symposium (KWEPSY) co-located with the 4th European Semantic Web Conference (ESWC2007)*, 2007.
17. A. Lait and B. Randell. An Assessment of Name Matching Algorithms. Technical report, Dept. of Computing Science, University of Newcastle upon Tyne, 1996.
18. P. F. Patel-Schneider, P. Hayes, and I. Horrocks. OWL Web Ontology Language: Semantics and Abstract Syntax. Recommendation, W3C, February 2004.
19. A. Polleres. From SPARQL to rules (and back). In *Proceedings of the 16th World Wide Web Conference (WWW2007)*, pages 787–796, Banff, Canada, May 2007.
20. J. Preece. *Designing Usability and Supporting Sociability*. John Wiley & Sons, Inc, 2000.
21. G. Tummarello, R. Delbru, and E. Oren. Sindice.com: Weaving the Open Linked Data. In *Proceedings of the International Semantic Web Conference 2007 (ISWC2007)*, volume 4825/2008, pages 552–565, Busan, Korea, November 2007.

Topic Extraction from Scientific Literature for Competency Management

Paul Buitelaar, Thomas Eigner

DFKI GmbH
Language Technology Lab & Competence Center Semantic Web
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
paulb@dfki.de

Abstract We describe an approach towards automatic, dynamic and time-critical support for competency management and expertise search through topic extraction from scientific publications. In the use case we present, we focus on the automatic extraction of scientific topics and technologies from publicly available publications using web sites like Google Scholar. We discuss an experiment for our own organization, DFKI, as example of a knowledge organization. The paper presents evaluation results over a sample of 48 DFKI researchers that responded to our request for a-posteriori evaluation of automatically extracted topics. The results of this evaluation are encouraging and provided us with useful feedback for further improving our methods. The extracted topics can be organized in an association network that can be used further to analyze how competencies are interconnected, thereby enabling also a better exchange of expertise and competence between researchers.

1 Introduction

Competency management, the identification and management of experts on and their knowledge in certain competency areas, is a growing area of research as knowledge has become a central factor in achieving commercial success. It is of fundamental importance for any organization to keep up-to-date with the competencies it covers, in the form of experts among its work force. Identification of experts will be based mostly on recruitment information, but this is not sufficient as competency coverage (competencies of interest to the organization) and structure (interconnections between competencies) change rapidly over time. The automatic identification of competency coverage and structure, e.g. from publications, is therefore of increasing importance, as this allows for a sustainable, dynamic and time-critical approach to competency management.

In this paper we present a pattern-based approach to the extraction of competencies in a knowledge-based research organization (scientific topics, technologies) from publicly available scientific publications. The core assumption of our approach is that such topics will not occur in random fashion across documents, but instead occur only

in specific scientific discourse contexts that can be precisely defined and used as patterns for topic extraction.

The remainder of the paper is structured as follows. In section 2 we describe related work in competency management and argue for an approach based on natural language processing and ontology modeling. We describe our specific approach to topic extraction for competency management in detail in section 3. The paper then continues with the description of an experiment that we performed on topic extraction for competency management in our own organization, DFKI. Finally, we conclude the paper with some conclusions that can be drawn from our research and ideas for future work that arise from these.

2 Related Work

Competency management is a growing area of knowledge management that is concerned with the "identification of skills, knowledge, behaviors, and capabilities needed to meet current and future personnel selection needs, in alignment with the differentiations in strategies and organizational priorities." [1] Our particular focus here is on aspects of competency management relating to the identification and management of knowledge about scientific topics and technologies, which is at the basis of competency management.

Most of the work on competency management has been focused on the development of methods for the identification, modeling, and analysis of skills and skills gaps and on training solutions to help remedy the latter. An important initial step in this process is the identification of skills and knowledge of interest, which is mostly done through interviews, surveys and manual analysis of existing competency models. Recently, ontology-based approaches have been proposed that aim at modeling the domain model of particular organization types (e.g. computer science, health-care) through formal ontologies, over which matchmaking services can be defined for bringing together skills and organization requirements (e.g. [2], [3]).

The development of formal ontologies for competency management is important, but there is an obvious need for automated methods in the construction and dynamic maintenance of such ontologies. Although some work has been done on developing automated methods for competency management through text and web mining (e.g. [4]) this is mostly restricted to the extraction of associative networks between people according to documents or other data they are associated with. Instead, for the purpose of automated and dynamic support of competency management a richer analysis of competencies and semantic relations between them is needed, as can be extracted from text through natural language processing.

3 Approach

Our approach towards the automatic construction and dynamic maintenance of ontologies for competency management is based on the extraction of relevant competen-

cies and semantic relations between them through a combination of linguistic patterns, statistical methods as used in information retrieval and machine learning and background knowledge if available.

Central to the approach as discussed in this paper is the use of domain-specific linguistic patterns for the extraction of potentially relevant competencies, such as scientific topics and technologies, from publicly available scientific publications. In this text type, topics and technologies will occur in the context of cue phrases such 'developed a tool for XY' or 'worked on methods for YZ', where XY, YZ are possibly relevant competencies that the authors of the scientific publication is or has been working on. Consider for instance the following excerpts from three scientific articles in chemistry:

…profile refinement method for nuclear and magnetic structures…
…continuum method for modeling surface tension…
…a screening method for the crystallization of macromolecules…

In all three cases a method is discussed for addressing a particular problem that can be interpreted as a competency topic: '*nuclear and magnetic structures*', '*modeling surface tension*', '*crystallization of macromolecules*'. The pattern that we can thus establish from these examples is as follows:

method for [TOPIC]

as in:

method for [nuclear and magnetic structures]
method for [modeling surface tension]
method for [(the) crystallization of macromolecules]

Other patterns that we manually identified in this way are:

approach for [TOPIC]
approaches for [TOPIC]
approach to [TOPIC]
approaches to [TOPIC]
methods for [TOPIC]
solutions for [TOPIC]
tools for [TOPIC]

We call these the 'context patterns', which as their name suggests provide the lexical context for the topic extraction. The topics themselves can be described by so-called 'topic patterns', which describe the linguistic structure of possibly relevant topics that can be found in the right context of the defined context patterns. Topic patterns are defined in terms of part-of-speech tags that indicate if a word is for instance a noun, verb, etc. For now, we define only one topic pattern that defines a topic as a noun (optional) followed by a sequence of zero or more adjectives followed by a

sequence of one or more nouns. Using the part-of-speech tag set for English of the Penn Treebank [5], this can be defined formally as follows - JJ indicates an adjective, NN a noun, NNS a plural noun:

(.?)((NN(S)? |JJ)*NN(S)?)*

The objective of our approach is to automatically identify the most relevant topics for a given researcher in the organization under consideration. To this end we download all papers by this researcher through Google Scholar run the context patterns over these papers and extract a window of 10 words to the right of each matching occurrence.

We call these extracted text segments the 'topic text', which may or may not contain a potentially relevant topic. To establish this, we first apply a part-of-speech tagger (TnT: [6]) to each text segment and sub-sequentially run the defined topic pattern over the output of this. Consider for instance the following examples of context pattern, extracted topic text in its right context, part-of-speech tagged version[1] and matched topic pattern (highlighted):

approach to
semantic tagging , using various corpora to derive relevant underspecified lexical
JJ NN *, VBG JJ NN TO VB JJ JJ JJ*
semantic tagging

solutions for
anaphoric expressions . Accordingly , the system consists of three major modules :
JJ NNS *. RB , DT NN VBZ IN CD JJ NNS :*
anaphoric expressions

tools for
ontology adaptation and for mapping different ontologies should be an
NN NN *CC IN VBG JJ NNS MD VB DT*
ontology adaptation

approach for
modeling similarity measures which tries to avoid the mentioned problems
JJ NN NNS *WDT VBZ TO VB DT VBN NNS*
modelling similarity measures

methods for
domain specific semantic lexicon construction that builds on the reuse
NN JJ JJ NN NN *WDT VBZ IN DT NN*
domain specific semantic lexicon construction

[1] Clarification of the part-of-speech tags used: CC: conjunction; DT, WDT: determiner; IN: preposition; MD: modal verb; RB: adverb; TO: to; VB, VBG, VBP, VBN, VBZ: verb

As can be observed from the examples above, mostly the topic to be extracted will be found directly at the beginning of the topic text. However, in some cases the topic will be found only later on in the topic text, e.g. in the following examples[2]:

approach to
be used in a lexical choice system , the model of
VB VBN IN DT JJ NN NN , DT NN IN
lexical choice system

approach for
introducing business process-oriented knowledge management , starting on the …
*VBG **NN** **JJ** **NN** **NN** , VBG IN DT …*
business process-oriented knowledge management

The topics that can be extracted in this way now need to be assigned a measure of relevance, for which we use the well-known TF/IDF score that is used in information retrieval to assign a weight to each index term relative to each document in the retrieval data set [7]. For our purposes we apply the same mechanism, but instead of assigning index terms to documents we assign extracted topics (i.e. 'terms') to individual researchers (i.e. 'documents') for which we downloaded and processed scientific publications. The TF/IDF measure we use for this is defined as follows:

$$D = \{d_1, d_2, \ldots, d_n\}$$

$$D_{freq>1}^{topic} = \{d_1, d_2, \ldots, d_n\} \text{where } \text{freq}_{d_i}^{topic} > 1 \text{ for } 1 \le i \le n$$

$$tf_d^{topic} = \frac{freq_d^{topic}}{freq_D^{topic}}$$

$$idf^{topic} = \frac{|D|}{|D_{freq>1}^{topic}|}$$

$$tfidf_d^{topic} = tf_d^{topic} * idf^{topic}$$

where D is a set of researchers and $freq_d^{topic}$ is the frequency of the topic for researcher d

The outcome of the whole process, after extraction and relevance scoring, is a ranked list of zero or more topics for each researcher for which we have access to publicly available scientific publications through Google Scholar.

[2] Observe that '*lexical choice system*' is a topic of relevance to NLP in natural language generation.

4 Experiment

To evaluate our methods we developed an experiment based on the methods discussed in the previous section, involving researchers from our own organization, DFKI. For all of these, we downloaded their scientific publications, extracted and ranked topics as explained above and then asked a randomly selected subset of this group to evaluate the topics assigned to them. Details of the data set used, the evaluation procedure, results obtained and discussion of results and evaluation procedure are provided in the following.

4.1 Data Set

The data set we used in this experiment consists of 3253 downloaded scientific publications for 199 researchers at DFKI. The scientific content of these publications are all concerned with computer science in general, but still varies significantly as we include researchers from all departments at DFKI[3] with a range of scientific work in natural language processing, information retrieval, knowledge management, business informatics, image processing, robotics, agent systems, etc.

The documents were downloaded by use of the Google API, in HTML format as provided by Google Scholar. The HTML content is generated automatically by Google from PDF, Postscript or other formats, which unfortunately contains a fair number of errors - among others the contraction of 'fi' in words like 'specification' (resulting in 'specication' instead), the contraction of separate words into nonsensical compositions such as 'stemmainlyfromtwo' and the appearance of strange character combinations such 'â€"'. Although such errors potentially introduce noise into the extraction we assume that the statistical relevance assignment will largely normalize this as such errors do not occur in any systematic way. Needless to say that this situation is however not ideal and that we are looking for ways to improve this aspect of the extraction process.

The document collection was used to extract topics as discussed above, which resulted first in the extraction of 7946 topic text segments by running the context patterns over the text sections of the HTML documents[4]. The extracted topic text segments (each up to 10 words long) were then part-of-speech tagged with TnT, after which we applied the defined topic pattern to extract one topic from each topic text[5]. Finally, to compute the weight of each topic for each researcher (a topic can be assigned to several researchers but potentially with different weights) and to assign a

[3] See http://www.dfki.de/web/welcome?set_language=en&cl=en for an overview of DFKI departments and the corresponding range in scientific topics addressed.

[4] For this purpose we stripped of HTML tags and removed page numbering, new-lines and dashes at end-of-line (to normalize for instance 'as-signed' to 'assigned').

[5] In theory it could also occur that no topic can be identified in a topic text, but this will almost never occur as the topic text will contain at least one noun (that matches the topic pattern as defined in section 3).

ranked list of topics to each researcher, we applied the relevance measure as discussed above to the set of extracted topics and researchers.

4.2 Evaluation and Results

Given the obtained ranked list of extracted topics, we were interested to know how accurate it was in describing the research interests of the researchers in question. We therefore randomly selected a subset of researchers from the 199 in total that we extracted topics for, including potentially also a number of researchers without assigned topics, e.g. due to sparse data in their case. This subset of researchers that we asked to evaluate their automatically extracted and assigned topics consisted of 85 researchers, out of which 48 submitted evaluation results.

The evaluation consisted of a generated list of extracted and ranked topics, for which the researcher in question was asked simply to accept or decline each of the topics. The evaluation process was completely web-based, using a web form as follows:

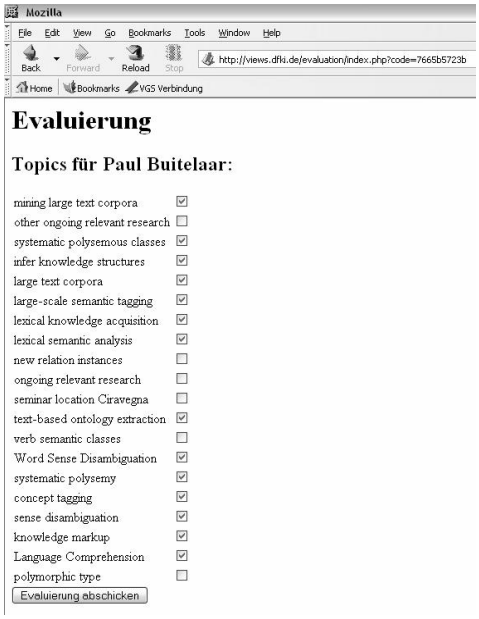


Figure 1: Web-form for evaluation of extracted topics

The evaluation for the 48 researchers that responded covered 851 extracted topics, out of which 380 were accepted as appropriate (44.65%). The following table provides a more detailed overview of this by distinguishing groups of researchers according to a level of how they judged their assigned topics correct ('Level of Correctness').

Level of Correctness	Number of Researchers
0-10%	*7*
11-20%	*1*
21-30%	*3*
31-40%	*9*
41-50%	*6*
51-60%	*9*
61-70%	*10*
71-80%	*3*
81-100%	*0*
	48

Table 1: Evaluation results

4.3 Discussion

Results of the evaluation vary strongly between researchers: almost half of them judge their assigned topics as more than 50% correct and 13 judge them more than 60% correct – on the other hand, 7 researchers are very critical of the topics extracted fro them (less than 10% correct) and slightly more than half judge their assigned topics less than 50% correct.

Additionally, in discussing evaluation results with some of the researchers involved we learned that it was sometimes difficult for them to decide on the appropriateness of an extracted topic, mainly because a topic may be appropriate in principle but it is: i) too specific or too general; ii) slightly spelled wrong; iii) occurs in capitalized form as well as in small letters; iv) not entirely appropriate for the researcher in question. We also learned that researchers would like to rank (or rather re-rank) extracted topics, although we did not explicitly tell them they were ranked in any order.

In summary, we take the evaluation results as a good basis for further work on topic extraction for competency management, in which we will address a number of the smaller and bigger issues that we learned out of the evaluation.

5 Applications

The overall application of the work presented here is management of competencies in knowledge organizations such as research institutes like DFKI. As mentioned we will therefore make the extracted topics available as ontology and corresponding knowledge base, on which further services can be defined and implemented such as expert finding and matching. For this purpose we need to organize the extracted topics further by extracting relations between topics and thus indirectly between researchers or groups of researchers working on these topics. We took a first step in this direction by analyzing the co-occurrence of positively judged topics (380 in total) from our evaluation set in the documents that they were extracted from. This resulted in a ranked listed of pairs of topics co-occurring more or less frequently. The following

table provides a sample of this (the top 15 co-occurring topics over the 1091 documents for the 48 researchers that responded to the evaluation task):

# of co-occurrences	Topic 1	Topic 2
1164	knowledge representation	knowledge base
796	information retrieval	knowledge base
676	question answering	knowledge base
528	question answering	information retrieval
524	knowledge representation	information retrieval
416	business process	business process modeling
416	knowledge representation	context information
384	information retrieval	context information
368	context information	knowledge base
364	information retrieval	sense disambiguation
360	business process	information retrieval
336	knowledge representation	question answering
336	linguistic processing	information retrieval
296	business process	knowledge base
292	knowledge markup	knowledge base

Table 2: Top-15 co-occurring topics

We can also visualize this as follows:

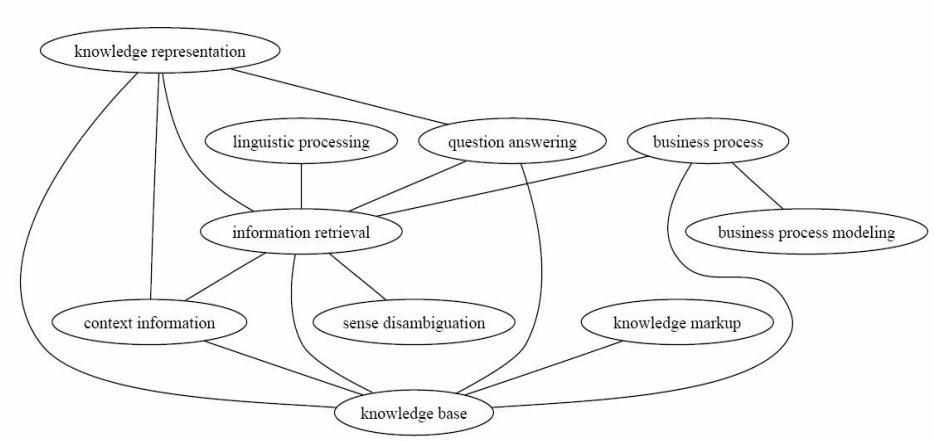


Figure 2: Association network between extracted topics (excerpt)

A different application that we are working on is to display the competencies of DFKI researchers in our web sites, e.g. by hyperlinking their names with an overview of competencies (scientific topics, technologies) that were either extracted automatically with the procedures discussed here or manually defined by the researchers themselves. For this purpose we integrate extracted topics into an individualized website on the DFKI intranet that allows each researcher to manage this as they see fit as follows:

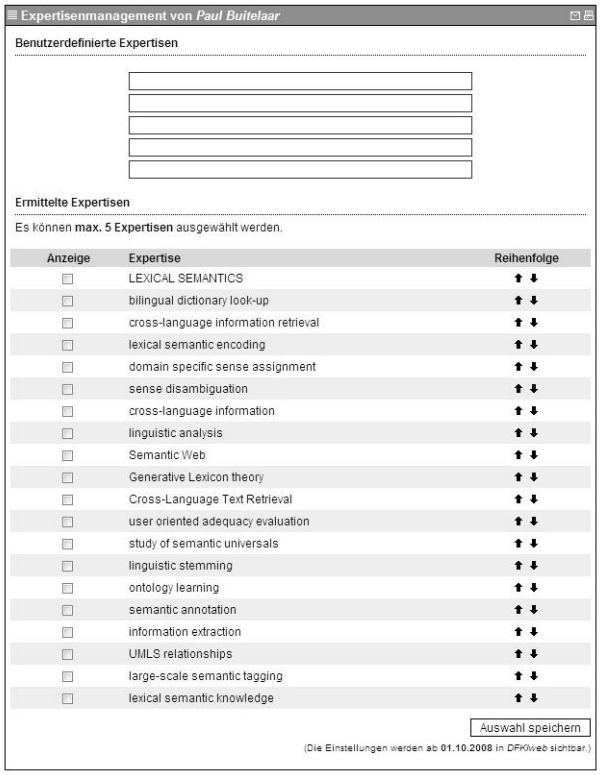


Figure 3: DFKI Intranet web-form for personalized expertise management

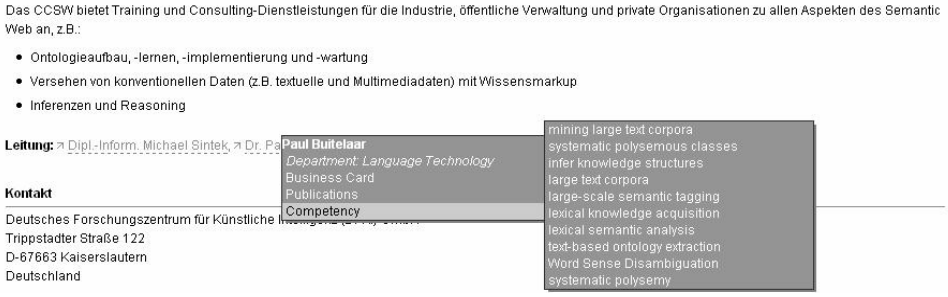


Figure 4: DFKI Intranet web application for expertise visualization

6 Conclusions and Future Work

In this paper we described an approach towards automatic, dynamic and time-critical support for competency management based on topic extraction from relevant text documents. In the use case we presented, we focus on the extraction of topics that represent competencies in scientific research and technology. Results obtained through an experiment on this for our own organization, DFKI, as example of a knowledge organization, are encouraging and provided us with useful feedback for improving our methods further. In current and future work we are therefore addressing some of the issues encountered during the evaluation process, in particular on improving the quality of the document collection, extending the coverage and precision of the topic and context patterns and further experimenting with the ranking scores we use.

Besides this we are currently extending the work on relation extraction between topics and (groups of) researchers as presented in an early stage in section 5, leading to methods for exporting extracted topics and relations as a shallow ontology with a corresponding knowledge base of associated researchers and documents that can be used to build further services such as semantic-level expert finding and matching.

Finally, we are currently preparing an extended evaluation that will include comparison with a baseline method on topic extraction, which does not use any specific context patterns as we defined and used them in our approach. For this purpose we are considering the use of TermExtractor[6], which enables the extraction of domain-relevant terms from a corresponding domain-specific document collection [8]. We consider the task of term extraction vs. topic extraction to be similar enough to justify this comparison.

Acknowledgements

This work was supported by the DFKI-internal project VIeWs4DFKI and in part by the Theseus-MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under grant number 01MQ07016.

We would like to thank our colleagues at DFKI for providing feedback on the evaluation of automatically extracted topics and two anonymous reviewers for providing constructive and very useful feedback for improving this paper as well as our methods and evaluation in general.

References

1. Draganidis, F. Mentzas, G.: Competency based management: A review of systems and approaches. Information Management and Computer Security 14(1), pp. 51–64 (2006)

[6] http://lcl2.uniroma1.it/termextractor/

2. Colucci, S., Di Noia, T., Di Sciascio, E., Donini, F., Piscitelli, G., Coppi, S.: Knowledge based approach to semantic composition of teams in an organization. In: SAC-05, ACM, New York, pp. 1314–1319 (2005)
3. Kunzmann, C., Schmidt, A.: Ontology-based Competence Management for Healthcare Training Planning: A Case Study. In: I-KNOW 2006, Graz & Special Issue of the Journal of Universal Computer Science (J.UCS), ISSN 0948-695X, pp. 143-150 (2006)
4. Zhu, J., Goncalves, A. L., Uren, V. S., Motta, E., Pacheco, R.: Mining Web Data for Competency Management. In: IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), pp. 94-100 (2005)
5. Marcus, M., Santorini, B., Marcinkiewicz, M.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19(2) (1993)
6. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: 6th ANLP Conference, Seattle, WA (2000)
7. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management 24/5, pp.515-523 (1988)
8. Sclano, F., Velardi, P.: TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. In: 3rd International Conference on Interoperability for Enterprise Software and Applications I-ESA 2007, Funchal, Madeira Island, Portugal (2007)

The Hoonoh Ontology for describing Trust Relationships in Information Seeking

Tom Heath[1] and Enrico Motta[2]

[1]Talis Information Limited
Knights Court, Solihull Parkway
Birmingham Business Park, B37 7YB, United Kingdom
firstname.surname@talis.com

[2]Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
initial.surname@open.ac.uk

Abstract. When seeking information through mechanisms such as word of mouth, people choose information sources and make trust judgments about these sources based on a range of factors, including the expertise of the source in relevant fields. In this paper we describe the Hoonoh Ontology, a vocabulary for describing these factors and publishing associated data on the Semantic Web. The ontology maps to existing vocabularies such as FOAF and SKOS, and when coupled with appropriate algorithms can be used to populate the Semantic Web with data to support expert finding initiatives.

1 Introduction

Numerous scenarios have been proposed in which expert-finding technology may be beneficial, such as human resources, electronic communications within communities and disaster response [3,4]. One similar scenario with day-to-day relevance for many people is information seeking via *word of mouth* recommendations. In such scenarios, where an individual encounters a problem or task for which their current knowledge is inadequate, they may engage in information-seeking in order to change their knowledge state [1]. Whilst the Web provides vast resources that may address the seeker's information need, "many information-gathering tasks are better handled by finding a referral to a human expert rather than by simply interacting with online information sources" (pp. 27) [9]. In typical word of mouth scenarios, the information seeker is faced with the task of finding the appropriate source who can help meet his or her information need.

In previous research [7] we investigated this issue, and found that the source selection process is influenced by five factors that determine the perceived trustworthiness of a source: *expertise*, *experience*, *impartiality*, *affinity* and *track record*. The first three of these factors (expertise, experience, impartiality)

represented a relationship between a person and a topic, whilst the latter two (affinity and track record) represented relationships between two people. For example, an individual may be perceived as an expert with respect to the topic of films, while two friends who have much in common may have a strong affinity. The reader is referred to [7] for fuller descriptions of each trust factor.

As reported in [8], we have developed algorithms that generate trust metrics based on these factors. These metrics can then be used in technical systems to help users identify experts and other people who may serve as relevant information sources. In the remainder of this paper we describe the Hoonoh Ontology[1], a vocabulary for describing these trust relationships relevant to the information seeking process.

2 The Hoonoh Ontology for Representing Computed Trust Relationships

The Hoonoh Ontology provides a vocabulary with which to represent computed trust metrics relevant to word of mouth information seeking. The ontology models person → topic and person → person relationships based on all five trust factors identified in our empirical research described above. Readers can view the ontology online at <http://hoonoh.com/ontology#>, while the following section describes the design of the ontology and related modeling decisions.

2.1 Modeling Trust Relationships

Nine classes are defined in total in the Hoonoh Ontology – eight of which relate to trust relationships and one to topics. Five of these are used to directly express trust relationships. `ExpertiseRelationship`, `ExperienceRelationship` and `ImpartialityRelationship` are subclasses of the `TopicalRelationship` class, and represent person → topic relationships. `AffinityRelationship` and `TrackRecordRelationship` are subclasses of the `InterpersonalRelationship` class and represent person → person relationships. `TopicalRelationship` and `InterpersonalRelationship` are not intended to be used to describe instance data but are provided simply as unifying superclasses, and are themselves subclasses of a unifying `Relationship` class.

Trust relationships are modeled in the Hoonoh ontology as instances of classes. This allows varying degrees of trust to be expressed by specifying numerical values as properties of these relationships. This is achieved using the `hoonoh:value`

1 The name “Hoonoh” is a play on the words “who” and “know”

property, which has an `rdfs:domain`[2] of `hoonoh:Relationship`[3] and an `rdfs:range` of `xsd:decimal`[4].

This modeling pattern was chosen for a number of reasons. Firstly, we found no evidence in our empirical work to suggest that trust was a binary relationship. Responses provided by participants in our study suggested that source selection decisions were rather subtle and nuanced, with trust relationships reflecting shades of grey rather than a binary 'trust/not trust' distinction.

Secondly, the algorithms we have developed to generate trust metrics based on our empirical work combine numerical data from a range of sources to compute a final metric for each factor. Inferring binary relations from such data would require the setting of an arbitrary numerical threshold at which to create a relationship, which would in turn limit the richness of relationships expressed in the ontology. Consequently, it was deemed preferable to expose numerical values for trust relationships and allow applications to interpret these as desired.

Our modeling approach contrasts somewhat with that adopted by [5], whose trust ontology allows trust relationships to be defined on a scale of 1-9, with each point in the scale having a dedicated property defined in the ontology. For example, 1 on the scale corresponds to the property 'distrustsAbsolutely', 5 to the property 'trustsNeutrally' and 9 to the property 'trustsAbsolutely'. While this approach does collapse into discrete values the notion of trust, which could be considered a continuous variable, it retains a reasonable degree of precision due to the use of a 9-point scale. However, we would argue that using a distinct property for each point in the scale adds complexity for those wishing to query the data with languages such as SPARQL [11].

2.2 Modeling People and Topics

The person from whom a relationship originates (the 'source') is identified using the `hoonoh:from` property, which has a domain of `hoonoh:Relationship` and a range of `foaf:Person`[5]. A class for people is not defined in the Hoonoh ontology; instead the Person class from the FOAF ontology [2] is reused to avoid duplication. For example, a relationship might exist between a person *A* and a topic *B*, or between a person *A* and another person *C*. In both cases the `hoonoh:from` property would be used to indicate the role of *A* in this relationship.

The topic to which experience, expertise and impartiality relationships relate is defined using the `hoonoh:toTopic` property, which has a domain of `hoonoh:Relationship` and a range of `hoonoh:Topic`, itself a subclass of the `Concept` class from the SKOS Vocabulary [10].

Figure 1 provides a schematic view of how an `ExpertiseRelationship` is modeled in the Hoonoh ontology.

2 prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3 prefix hoonoh: <http://hoonoh.com/ontology#>
4 prefix xsd: <http://www.w3.org/2001/XMLSchema#>
5 prefix foaf: <http://xmlns.com/foaf/0.1/>

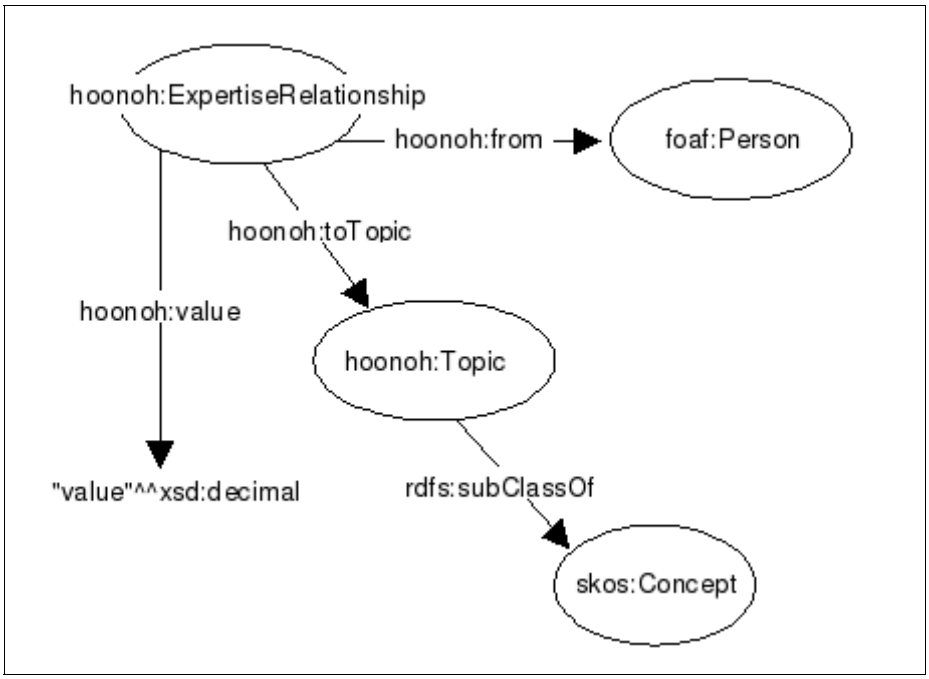


Fig. 1. Schematic diagram showing how expertise relationships are modeled in the Hoonoh ontology

In contrast to descriptions of experience, expertise and impartiality relationships, the description of affinity and track record relationships is completed by use of the `hoonoh:toPerson` property which defines the individual to whom the relationship refers. This property has a domain of `hoonoh:Relationship` and a range of `foaf:Person`, as shown in Figure 2.

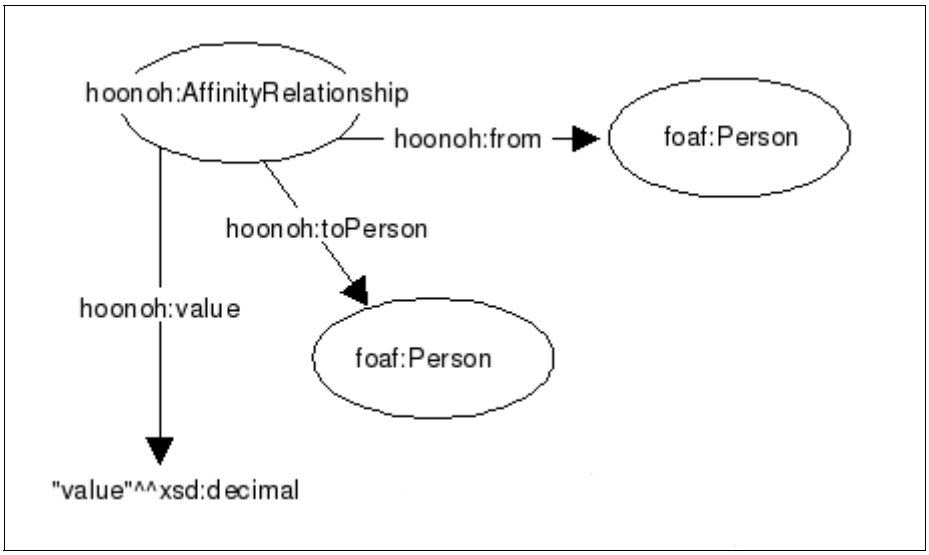


Fig. 2. Schematic diagram showing how affinity relationships are modeled in the Hoonoh ontology

To complement the schematic views, Code Fragment 1 and Code Fragment 2 below show examples of how an `ExpertiseRelationship` and an `AffinityRelationship` can be modeled using the Hoonoh ontology[6].

6 URIs shown in the http://hoonoh.com/ namespace, and values of foaf:mbox_sha1sum are deliberately shortened due to formatting limitations.

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
   ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
   xmlns:hoonoh="http://hoonoh.com/ontology#"
   xmlns:foaf="http://xmlns.com/foaf/0.1/"
   xml:base="http://hoonoh.com/">

<hoonoh:ExpertiseRelationship

rdf:about="relationships/expertise/abc123/example">
 <hoonoh:from rdf:resource="people/abc123"/>
 <hoonoh:toTopic rdf:resource="topics/example"/>
 <hoonoh:value
     rdf:datatype=
      "http://www.w3.org/2001/XMLSchema#decimal">
      0.7292
 </hoonoh:value>
</hoonoh:ExpertiseRelationship>

<foaf:Person rdf:about="people/abc123">
 <foaf:mbox_sha1sum>abc123</foaf:mbox_sha1sum>
</foaf:Person>

<hoonoh:Topic rdf:about="topics/example">
 <rdfs:label>example</rdfs:label>
</hoonoh:Topic>

</rdf:RDF>
```

Code Fragment 1. An example Expertise relationship described using the Hoonoh ontology

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
   ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns:hoonoh="http://hoonoh.com/ontology#"
   xmlns:foaf="http://xmlns.com/foaf/0.1/"
   xml:base="http://hoonoh.com/">

<hoonoh:AffinityRelationship
     rdf:about="relationships/affinity/abc123/xyz789">
 <hoonoh:from rdf:resource="people/abc123"/>
 <hoonoh:toPerson rdf:resource="people/xyz789"/>
 <hoonoh:value
     rdf:datatype=
      "http://www.w3.org/2001/XMLSchema#decimal">
      0.8500
 </hoonoh:value>
</hoonoh:AffinityRelationship>

<foaf:Person rdf:about="people/abc123">
 <foaf:mbox_sha1sum>abc123</foaf:mbox_sha1sum>
</foaf:Person>

<foaf:Person rdf:about="people/xyz789">
 <foaf:mbox_sha1sum>xyz789</foaf:mbox_sha1sum>
</foaf:Person>

</rdf:RDF>
```

Code Fragment 2. An example Affinity relationship described using the Hoonoh ontology

3 The Hoonoh Ontology in Practice

Having developed the Hoonoh Ontology, we have used algorithms similar to those described in [8] to generate trust metrics based on data from Revyu.com [6] and the *del.icio.us* social bookmarking service[7]. These metrics are described in RDF according to the Hoonoh Ontology and published online at Hoonoh.com[8], as crawlable RDF and via a SPARQL endpoint[9], to enable reuse in other applications.

7 http://del.icio.us/
8 http://hoonoh.com/
9 http://hoonoh.com/sparql

It is worth noting that both the Hoonoh ontology and the Hoonoh triplestore are oriented specifically towards describing and storing trust relationship data about individuals who have trust relationships generated by the algorithms, not generic information such as a names or home page addresses. The Hoonoh ontology and triplestore provide the necessary hooks with which these trust metrics can be merged with other data sources on the Semantic Web, such as an individual's FOAF file. In this way, trust metrics can be published on the open Web, whilst detailed personal information can remain under the control of the individual.

Building on the trust data in the Hoonoh triplestore we have implemented the Hoonoh.com social search engine[10], which enables users to search for topics of interest, and receive suggestions of trusted information sources from among members of their social network. Results (i.e. members of the user's social network who may have relevant information) are ranked according to the trust metrics described with the Hoonoh ontology. The result is an application that allows people to search for information online using the same criteria for selection of information sources that they may use when seeking information offline.

4 Conclusions

In this paper we have presented the Hoonoh Ontology for describing trust relationships in the context of word of mouth information seeking. The ontology itself has been developed based on empirical research in the field and as such provides a domain model with high ecological validity. While the ontology is not specific to describing individuals' expertise, it does enable these relationships to be expressed, thereby making it suitable for use in expert-finding applications. Furthermore, it also provides the means to model a number of other relationships that, while they may not be isomorphic to expertise, remain highly relevant to applications and services in this domain.

References

[1] Belkin, N. J. Helping People Find What They Don't Know. Communications of the ACM, 43, 58-61. 2000.

[2] Brickley, D. and Miller, L. FOAF Vocabulary Specification 0.9. http://xmlns.com/foaf/0.1/ (accessed 10th August 2008)

[3] ExpertFinder Initiative. http://expertfinder.info/ (accessed 10th August 2008)

10 http://hoonoh.com/

[4] ExpertFinder User Cases. http://wiki.foaf-project.org/ExpertFinderUseCases (accessed 10th August 2008)

[5] Golbeck, J., Parsia, B. and Hendler, J. Trust Networks on the Semantic Web. In Proceedings of the Workshop on Cooperative Information Agents (CIA2003), Helsinki, Finland. 2003.

[6] Heath, T. and Motta, E. (2008) Ease of Interaction plus Ease of Integration: Combining Web2.0 and the Semantic Web in a Reviewing Site. Journal of Web Semantics, 6 (1) (Special Issue on Web2.0 and the Semantic Web).

[7] Heath, T., Motta, E., and Petre, M. Person to Person Trust Factors in Word of Mouth Recommendation. In Proceedings of the CHI2006 Workshop on Reinventing Trust, Collaboration, and Compliance in Social Systems (Reinvent06), Montréal, Canada. 2006.

[8] Heath, T., Motta, E. and Petre, M. Computing Word-of-Mouth Trust Relationships in Social Networks from Semantic Web and Web2.0 Data Sources. In Proceedings of the Workshop on Bridging the Gap between Semantic Web and Web 2.0, 4th European Semantic Web Conference (ESWC2007), Innsbruck, Austria. 2007.

[9] Kautz, H., Selman, B. and Shah, M. The Hidden Web. AI Magazine, Summer. 1997.

[10] Miles, A. and Bechhofer, S. SKOS Simple Knowledge Organization System Reference. http://www.w3.org/TR/skos-reference/ (accessed 10th August 2008).

[11] Prud'hommeaux, E. and Seaborne, A. SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/ (accessed 10th August 2008).

The 7th International Semantic Web Conference

October 26 – 30, 2008

Congress Center, Karlsruhe, Germany