Extracting Semantic User Networks From Informal Communication Exchanges

A.L. Gentile, V. Lanfranchi, S. Mazumdar, and F. Ciravegna

Department of Computer Science
University of Sheffield
Sheffield, United Kingdom
{a.l.gentile,v.lanfranchi,s.mazumdar,f.ciravegna}@dcs.shef.ac.uk

Abstract. Nowadays communication exchanges are an integral and time consuming part of people's job, especially for the so called *knowledge workers*. Contents discussed during meetings, instant messaging exchanges, email exchanges therefore constitute a potential source of knowledge within an organisation, which is only shared with those immediately involved in the particular communication act. This poses a knowledge management issue, as this kind of contents become "buried knowledge". This work uses semantic technologies to extract buried knowledge, enabling expertise finding and topic trends spotting. Specifically we claim it is possible to automatically model people's expertise by monitoring informal communication exchanges (email) and semantically annotating their content to derive dynamic user profiles. Profiles are then used to calculate similarity between people and plot *semantic knowledge-based networks*. The major contribution and novelty of this work is the exploitation of semantic concepts captured from informal content to build a semantic network which reflects people expertise rather than capturing social interactions. We validate the approach using contents from a research group internal mailing list, using email exchanges within the group collected over a ten months period.

1 Introduction

Email is a common tool for quick exchange of information between individuals and within groups, especially in formal organisations [10], via the use of official or unofficial mailing lists. Mailing lists exchanges are often used to reach a wider audience that includes both the initiator's personal networks and other individuals with shared interests [31] and who may be potential sources of expertise. This poses a knowledge management issue, as the emails' knowledge content is not shared with the whole organisation but only with those included in the recipient list, thus implicitly creating and/or reinforcing the existence of dynamic communtities inside organisations. Whilst this is positive as it increases flexibility and innovation, the drawback is that knowledge remains implicit or not shared with the rest of the organisation, becoming "buried knowledge" [32]. Moreover as recognised by [25] the lines between inter-communication on professional and social levels are increasingly blurred: people tend to share more aspects of their social life with co-workers and these exchanges often lead to establishing professional coooperations or sharing topics of interests. Extracting information from emails could prove useful in a knowledge management perspective, as it would provide means to build social networks, determine experts and communities of practice, taking into account not only professional content but also social and emerging topics, that may highlight emerging cooperations, interests etc. As proved by [14] email traffic can be analysed and used to identify COINS (Collaborative Innovation Networks), groups of individuals inside organisations that are self-motivated and work together on a new idea. In this paper we propose an approach to automatically model people's expertise and dynamic communities interests by monitoring informal communication exchanges. The content of communication exchanges is semantically annotated and used to derive user profiles. Profiles are then used to calculate similarity between people and plot *semantic knowledge-based networks*, highlighting groups of users with shared knowledge or with complementary knowledge. The main novelty concerns the profile generation; with respect to the state of the art in the Social Network Analysis, where profiles are mostly based (i) on information declared by users in their static profiles, (ii) on rates of communication exchanges between users and (iii) on the morphology of the social graph, our work proposes a profile generation which is based on semantic concepts extracted from user generated content; these profiles have the advantage of being both dynamic, as they are created from user generated content and semantic, in the sense that unique and meaningful concepts are extracted. To confirm its quality and validity we experimented the proposed approach on a dataset consisting of a research group internal mailing list exchanges (as in [13]), extracting profiles with different degrees of semantics. The similarity between users has then been calculated for every type of profile and compared against users subjective similarity judgements, to understand if increasing the level of semantics in the user profiles increases the accuracy of similarity, therefore increasing the potential for exploiting the technique for different tasks, such as expert finding, knowledge sharing or replicated expertise detection within an organisation. To demonstrate the usefulness of the approach, this has been integrated into a knowledge management system aimed at capturing and sharing knowledge exchanged during informal communication exchanges (meetings, chats, etc.): this has provided clear scenarios of usage that will be evaluated in the coming months to understand whether semantic dynamic user profiling is beneficial to increase efficiency and efficacy in knowledge management.

The paper is structured as follows. Section 2 presents a review of the state of the art in knowledge capture from informal communication exchanges, on semantic user profiling and on measures for user similarity. Section 3 describes the proposed approach to automatically model people's expertise profiles and calculate similarity between them. Section 4 presents the experiments used to extract and evaluate user profiles and similarities, whilst section 5 introduces the applications of the approach and possible scenarios of use. Section 6 discusses the results and the next stages of our research.

2 State of the Art

2.1 Knowledge Capture from Informal Communication Exchanges

Capturing, representing and sharing knowledge from informal communication exchanges has been a topic of research in the knowledge management and in the information seeking behaviour communities for many years, as this type of knowledge is tacit, often very specilized and precise and is not shared with anyone else than the immediate recipient of the informal communication exchange. Previous researches on knowledge workers information seeking behaviours proved how engineers spend 40-66% of their time sharing information [18, 30]. Different types of communication exchange can be recognised at different levels of formality, with structured planning meetings or corporate mailing lists at one side of the spectrum and informal chats at the coffee machine or chats over internet messaging programs at the other. Independently from the degree of formality all these communication exchanges contain invaluable knowledge that is often buried [32]. Different techniques have been explored in previous works to extract, represent and share this buried knowledge, often focusing on one specific type of communication exchange, such as emails, meeting recordings etc. In this work we focus on email exchanges. Email content, and addressee and recipient, often provide clues about the interests and expertise of participants [5] and they are used as a source for automatic expertise identification and knowledge elicitation [7, 31]. The main techniques adopted to extract information and build social networks from emails are usually quantitative data analysis, such as frequency of exchanges between individuals, and data mining over the email content.

Exchange Frequency In the panorama of work on extracting social networks from email, the frequency of email exchange has been widely used as the main indicator of relevance of a connection. In some cases the effort is on determining frequency thresholds [33, 11, 3, 9], while in others time-dependent threshold conditions are defined to detect dynamic networks [6, 19]. Diesner et al. [10] construct a social network via weighted edges over a classical dataset, the *Enron* corpus[1], a large set of email messages made public during the legal investigation of the Enron corporation. They reported the emergence of communication subgroups with unusually high email exchange in the period prior to the company becoming insolvent in 2001, when email was a key tool for obtaining information especially across formal, inter-organisational boundaries. Diesner et al. [10] also observed that variations in patterns of email usage were influenced by knowledge about and reputation of, in addition to, formal roles within the organisation. Our approach differs from the above cited works as we choose to not take into account the frequency of individual communication exchanges but to perform content-based analysis of mailing list archives, where the emphasis is not on individual recipients of an email but on reaching a wider base of colleagues to find experts that could answer very specific questions.

[1] `http://www.cs.cmu.edu/~enron`

Content-Based Analysis Email content analysis has been used for different purposes: determining expertise [31], analysing the relations between content and people involved in email exchanges [5, 17, 25, 39], or simply extracting useful information about names, addresses, phone numbers [21]. Our approach takes inspiration from Schwartz et al. [31] in trying to derive expertise and common interests within communities from email exchange but moves on from the keyword-based extraction approach, to consider McCallum et al. [25] contribution in applying Machine Learning (ML) and Natural Language Processing (NLP) to retrieve the rich knowledge content of the information exchanged in automatically gathered social networks, and better interpret the attributes of nodes and the types of relationships between them.

Laclavík et al. [21] observe that enterprise users largely exploit emails to communicate, collaborate and carry out business tasks. They also adopt a cont-based approach and exploit pattern-based Information Extraction techniques to analyse enterprise email communications, and exploit the data obtained to create social networks. The test sets (one in English containing 28 emails, and a second in Spanish with 50 emails) consist of mainly formal emails exchanged between different enterprises. The results obtained indicate that emails are a valid means for obtaining information across formal, inter-organisational boundaries. Lin et al. [24, 23]propose a framework for social-oriented Knowledge Management within an organization. They exploit the content of emails for corporate search and for providing expert finding and social network facilities. The focus of these work is rather on the framework description in terms of provided functionalities than on the description of how the content is processed and how the user profiles are generated. The work we present in this paper, on the other hand, makes use of a test set containing informal email exchanges from an internal mailing list for an academic research group, for a pilot, exploratory set of experiments. We adopt multiple approaches for extracting information at different levels of semantic granularity to aid the understanding of the content of the conversations carried out via email, and depict the variety of topics discussed using this communication medium and ultimately derive effective user profiles.

2.2 Semantic User Profiles

Using semantic technologies to derive user profiles has been a topic of research in recent years within different communities, especially in the field of Information Retrieval, with the objective of providing customized search results and in the Recommender Systems community, with the aim of generating effective customized suggestions to the users. In the IR community the focus is on building a user profile which reflects the user interests more than the user expertise, to customize search results. Daoud et al. [8] represent semantic user profiles as graphs of concepts derived from a pre-defined ontology and represent documents as vector of concepts of the same ontology. Then profiles are exploited to determine the personalized result ranking, using a graph-based document ranking model. In the field of Recommender Systems the goal is improving the accuracy recommendations. Abel et al. [1] build semantic user profiles for news recommendations: they enrich users' Twitter activities with semantics extracted from news articles and then use these profiles to suggest users articles to read.

The input for the semantic user profile generation can be gathered in different ways: Kramar [20] e.g. proposes to monitor user activities on the web end extracting metadata (tags, keywords, named entities) from the visited documents. Another direction is the usage of user generated content to extract interests and knowledge levels of users. The most similar work to ours in this direction is the one proposed by Abel et al. [2] who use Twitter posts to generate User Models. Semantic user profiles are extracted from the messages people post on Twitter; in particular they generate three types of profiles: hashtag-based, topic-based or entity-based profiles. The main focus of their work is how to extract semantic from short text like tweets rather than extensively comparing the different types profiles. Our work also propose to generate three types of profiles, with increasing level of semantics (keyword, named entities and concepts), but we also investigate how increasing the levels of semantic in the user profile improves the quality of the profile for a certain task; specifically we show how a more semantic profile better reflects the user perceived similarity with other users in the community.

2.3 Measures for User Similarity

Calculating similarity between user is a key research question in many fields. In Social Networks the similarity between two users is usually a binary function, indicating the "friendship" of two users who are either connected or not. Non-binary similarities have also been proposed [35]. The relationship between two users can be explicit, as stated by the two users, or predicted by means of automatic techniques. Typical features which are used to infer the similarity are attributes from the user's profile like geographic location, age, interests [27]. Social connections already present in the graph are also used as features to predict new possible connections. Other commonly used features (for example in Facebook) are interaction-counters top-friend, and picture graphs. Also spatio-temporal data pertaining to an individuals trajectories has been exploited to geographically mine the similarity between users based on their location histories [22, 37].

In our study we take inspiration from the way content-based recommender systems create user profiles as vectors of defining keywords [4, 28]; we build user profiles by exploiting the content of emails belonging to each user and we represent them as vectors. The novelty of our technique is using semantic concepts as feature representation rather than keywords. With respect to existing techniques for user similarities in Social Networks, the novelty of our technique is exploiting user generated content to build the feature space, rather than exploiting static features from user profiles. We calculate similarity between users within a network but we use dynamic and semantic features rather that static user-defined ones. The main advantage is that we capture the dynamicity and evolutionary nature of the user interaction.

3 Semantic Network Extraction from Informal Mail Exchanges

The proposed approach models people's expertise and dynamic communities interests by analysing informal communication exchanges. The approach performs the generation of content-based user profiles by analysing user generated contents. User profiles are then used to derive similarity between users. The value of similarity can then be exploited to plot semantic network between users, which will reflect the similarity on the basis of shared knowledge. The following sections will discuss the profile generation and similarity derivation in details.

3.1 Building User Profiles

User profiles are built by extracting information from email content using three techniques, with varying degrees of semantics. This is done to ascertain the quality of user profiles and their suitability to model people's interests and expertise.
Keyword-based profile Each email e_i in the collection E is reduced to a Bag of Keywords representation, such as $e_i = \{k_1, \ldots, k_n\}$. Each user keyword-based profile consists of Bag of Keywords, extracted from their sent emails.
Entity-based profile Each email e_i in the collection E is reduced to a Bag of Entities representation, such as $e_i = \{ne_1, \ldots, ne_k\}$. Entities are elements in text which are recognized as belonging to a set of predefined categories (classical categories are persons, locations, organizations, but more fine grained classification is typically used). Each user entity-based profile consists of Bag of Entities, extracted from their sent emails.
Concept-based profile. Each email e_i in the collection E is reduced to a Bag of Concepts representation, such as $e_i = \{c_1, \ldots, c_n\}$. Concepts are elements in text which are identified as unique objects and linked to an entry in a reference Knowledge Base (Wikipedia in this case). Each user concept-based profile consists of Bag of Concepts, extracted from their sent emails.

Implementation details The keyword extraction process has been performed using Java Automatic Term Recognition Toolkit (JATR v1.0[2]). JATR implements a voting mechanism to combine the results from different methods for terminology recognition (dealing with single- and multi-word term recognition) into an integrated output, improving results of integrated methods taken separately [38].

The Named Entity extraction has been performed using the Open Calais web service[3]. Open Calais is an ontology-based service which returns extraction results in RDF. Together with named entity recognition, the service performs instance recognition (concept identification) and facts extraction. For the purposes of this work we only exploited the Named Entities returned by OpenCalais.

[2] http://staffwww.dcs.shef.ac.uk/people/Z.Zhang/resources/tools/jatr_v1.0.zip

[3] http://www.opencalais.com/

The Concept extraction process has been performed using the Wikify web service [26]. Wikify uses a machine-learning approach to annotate Wikipedia concepts within unstructured text. The disambiguation procedure uses three features: a priori probability of each Wikipedia concept, weighted context terms in the target text and a measure of the goodness of the context. The context terms weights are calculated by using the average semantic relatedness with all other context terms. The measure of goodness of the context reflects its homogeneity and it is used to dynamically balance a priori probability and context term weights, instead of using a fixed heuristic. The candidate terms are generated by gathering all n-grams in the text and discarding those below a low threshold (to discard non-sense phrases and stopwords). Experiments show that the method perform as well on non-Wikipedia texts as on Wikipedia ones, with an F-measure of 75% on non-Wikipedia texts.

3.2 Deriving People Similarity

The obtained user profiles are then used to calculate the similarity strength between users, measured on a [0,1] range. Similarity values reflect the amount of knowledge shared between two users. When used to plot a network of users similarity values can be useful to identify central users, small communities with shared knowledge (users with higher values of similarity among each other).

Following [15] similarity score is calculated using Jaccards index. The Jaccard similarity coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. Sample sets for our user similarity are concepts (or keywords or Named Entities respectively) in each user profile. Moreover the similarity calculated over semantic user profiles is compared with the same measure calculated over the keyword based profiles and the named entity based profiles, to prove that increasing the semantic value of a profile increases its quality and suitability for modelling people's expertise. The semantic profiles amongst the others, better mimic the users' perceived similarities between each other. Results for similarities calculated over the three different types of profiles are shown in section 4.

4 Experiments

The aim of the experiments was to validate the hypothesis that increasing the level of semantic in the user profile generation improves the quality of profiles. For such purpose we used the task of inferring similarities between users and assessed the correlation with human judgment. The corpus used for analysis and knowledge content extraction is an internal mailing list of the OAK Group in the Computer Science Department of the University of Sheffield[4]. The mailing list is used for quick exchange of information within the group on both professional and social topics. We use the same corpus as [13], but with a broader period

[4] http://oak.dcs.shef.ac.uk

coverage: we selected all emails sent to the mailing list in the ten month period from July 2010 to May 2011, totalling 1001 emails. We will refer to this corpus as *mlDataset*. For each email we extracted the subject and the email body. The number of users in the mailing list is 40; 25 of them are active users (users sending email to mailing list). The average message length is 420 characters. The average message length per user is shown in table 1, together with number of concepts in each user profile. Table 1 reports statistic about all 25 active users, even if only 15 of those 25 participated to the evaluation exercise.

ID	AvgMsg	n. sent email	Conc	ID	AvgMsg	n. sent email	Conc
1	445	62	42	16	612	121	71
3	453	209	90	18	170	27	35
5	1192	9	10	21	290	24	25
6	543	5	5	22	155	27	14
7	489	13	11	23	345	53	39
8	330	9	14	24	271	10	14
9	462	74	67	25	399	65	79
10	237	37	23	27	236	36	33
11	282	90	80	28	523	20	23
12	841	23	40	33	102	8	5
13	766	12	12	36	227	30	15
14	338	17	18	40	224	3	3
15	516	17	22				

Table 1. Corpus statistics for *mlDataset*. Column ID contains an anonymous identifier for the users. Column AvgMsg contains the average message length for that user, expressed in number of characters. Column Conc contains the number of concepts in the user profile.

We compared user similarity obtained with the three different profile types (section 3) against the users perceived similarity. Pearson correlation has been calculated for user judgments compared against the automatic generated similarities using Keyword based profiles, Named Entity based profiles and Concepts based profiles.

The evaluation was conducted as a paper-based exercise in which the participants were asked to fill in a table containing a list of people within the working group and rate their perceived similarity on a scale from 1 to 10, with *1 = not similar at all* and *10 = absolutely similar*. If the user was not known they were instructed to leave the rating blank. The participants were asked to score the similarity in terms of topics or interest shared with the other person, both from a professional and social point of view (e.g. hobbies or other things which emerge within the working time) to concentrate the user thoughts towards ”general” similarity without thinking about what specific type of similarity they share. The exercise was repeated twice for a small subset of users, with the second one seven days after the first one. The inter-annotator agreement was calculated by comparing for each user his/her perceived similarities about other participants and similarities perceived by all the rest of participants for that particular user. Inter-annotator agreement is shown in the last column of table 2.

A total of 15 users took part in the evaluation exercise, all providing valid questionnaires.

ID	K		NEs		Conc		Agr
	C	S	C	S	C	S	C
14	0.55	0	0.41	0.04	0.68	0	0.91
7	0.48	0.02	0.39	0.06	0.58	0	0.87
28	0.5	0.01	0.41	0.04	0.57	0	0.89
10	0.47	0.02	0.39	0.05	0.57	0	0.94
27	0.32	0.11	0.29	0.16	0.48	0.02	0.92
21	0.34	0.11	0.42	0.04	0.42	0.04	0.91
1	0.35	0.02	0.32	0.11	0.42	0.04	0.94
3	0.3	0.14	0.31	0.14	0.38	0.06	0.86
9	0.28	0.18	0.36	0.07	0.38	0.06	0.9
18	0.5	0.01	0.5	0.01	0.36	0.07	0.87
8	0.17	0.53	0.19	0.48	0.35	0.18	0.82
11	0.59	0	0.42	0.04	0.34	0.1	0.83
25	0.25	0.22	0.33	0.11	0.3	0.14	0.73
23	0.21	0.32	0.33	0.1	0.19	0.36	0.86

Table 2. Correlation (C) of similarity with user judgment at (S) significance level, obtained using Keyword based profiles (K), Named Entity based profile (NEs), Concept based profiles (Conc). Column (Agr) reports Inter-annotator agreement for each user at significance < 0.001

Table 2 shows the Pearson correlation between automatically generated similarity with user judgment. For the three types of user profiles, Keyword based profiles (K), Named Entity based profile (NEs) and Concept based profiles (Conc), the table shows the correlation value (C) and the respective significance level (S). The inter-annotator agreement for each user reported in column Agr, has been calculated with Pearson correlation at significance < 0.001. Results are presented in descending order of correlation for similarities over concept based profiles. Figures show that the correlation almost always improves by the usage of Concept based profiles over Keyword based profiles, except for three users (18, 11, 23). Moreover the significance level for the correlation on concept based similarity is lower than the one on keyword based similarity (except for user 23). For the three users not confirming the trend, the inter-annotator agreement average (0.83) is lower than the general average (0.88).

5 Applications and Scenarios of Usage

The approach presented in this paper allows to automatically extract content and user similarity from an email corpus to build networks which reflects people expertise rather than capturing users' social interactions. This approach could prove particularly useful for knowledge management, in particular for tasks such as expertise finding, trend spotting and identification of dynamic communities inside an organisation. In order to prove its usefulness for knowledge management the approach has been integrated into a knowledge management system aimed at capturing and sharing knowledge exchanged during informal communication exchanges (meetings, chats, etc.): this has provided clear scenarios of usage, which will be helpful to understand whether semantic dynamic user profiling increases efficiency and efficacy in expert finding and trend spotting tasks. The knowledge management framework adopts semantic user profiling to capture content from numerous informal communication exchanges such as emails, meeting recordings and minutes etc.; these are then visualised using SimNET (Similarity and Network Exploration Tool), a tool for exploring and searching over socio-centric content, that interactively displays content and users networks as part of the searching and browsing capabilities provided by the knowledge management system. Section 5.1 introduces SimNET, whilst section 5.2 introduces two scenarios of usage of the knowledge management system within a large organisation. The scenarios are presented to highlight the capabilities and usefulness of the approach; an evaluation of the expert finding and trend analysis tasks is scheduled for the coming months.

5.1 Semantic Network Visualisation

SimNET is a dynamic and real-time filtering interface that offers multiple visualisations for knowledge extracted from user generated content, using classical techniques such as node-Link diagrams [12, 29], force-directed layouts [16],[34], radial layouts [36] (as shown in Figure 1) and tag clouds (detail of a tag cloud in Figure 3). SimNET has been built as a flexible visualisation tool to explore socio-centric content using the most suitable visualisation for the undertaken task. Radial visualization is provided to focus on the interactions between users, while force-directed layout is provided to support by-topic knowledge exploration and trends observation. SimNET has two main interaction paradigms - email visualisation and similarity visualisation. The users are initially presented with an interface that visualises email interactions and a tag cloud describing all the concepts. The users can then choose between a radial or a force-directed layout according to the task and can use filtering and searching widgets to focus the exploration of the dataset. For example, when clicking on concepts in the tag cloud relevant emails are highlighted in the radial layout and vice versa. Users can also select to visualise similarities by clicking on 'Plot Similarity'. The radial graph is updated to show the similarities as edges between nodes. These edges are colour coded to signify the similarity among the connecting nodes. The interface provides the users with a *similarity slider*, which can be dragged to set a similarity threshold for showing edges and a *date range selection bar*, to explore the network evolution over time.

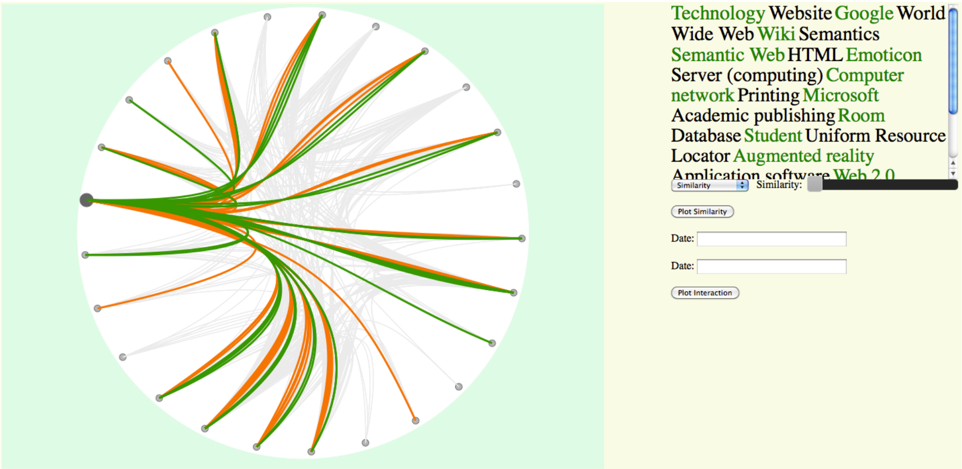
Technology Website Google World Wide Web Wiki Semantics Semantic Web HTML Emoticon Server (computing) Computer network Printing Microsoft Academic publishing Room Database Student Uniform Resource Locator Augmented reality Application software Web 2.0
Similarity
Similarity:
Plot Similarity
Date:
Date:
Plot Interaction

Fig. 1. SimNET, Similarity and Network Exploration Tool

Providing visualisation capabilities is critical for making sense of user profiles, topics and similarities as it allows users to access and manipulate knowledge.

5.2 Scenarios of use

Expertise Finding Our system supports the task of finding an expert for a given topic in multiple ways. For example a user may browse the tag cloud or perform a specific query. The system will then plot the results using a force-directed or a radial layout and the user will be able to interact with the visualisation to explore it and highlight users that are experts on that specific topic. This is very important as it addresses the long tail of information, allowing to discover expertise for topics that are not well known or for which not many people are knowledgeable. It is even more important when applied to large and dynamic organisation where the number of users is very high and it is very likely that they do not know each other and they are not aware of who are the current experts on certain topics, who are similar users to involve in a certain activity. Having a system that highlights the long-tail of information allows sharing knowledge in a more efficient manner; for example if a user working in a big organization is looking for information of a common and well-known topic within her/his department almost everyone in her/his group will be able to answer her request, whilst if looking for more obscure information that is interdisciplinary, people in the same department may not know the answer. In such a case a system that allows to discover expertise tailoring the short and long tail of information is invaluable as it quickly highlights people in the organisations for quick help.

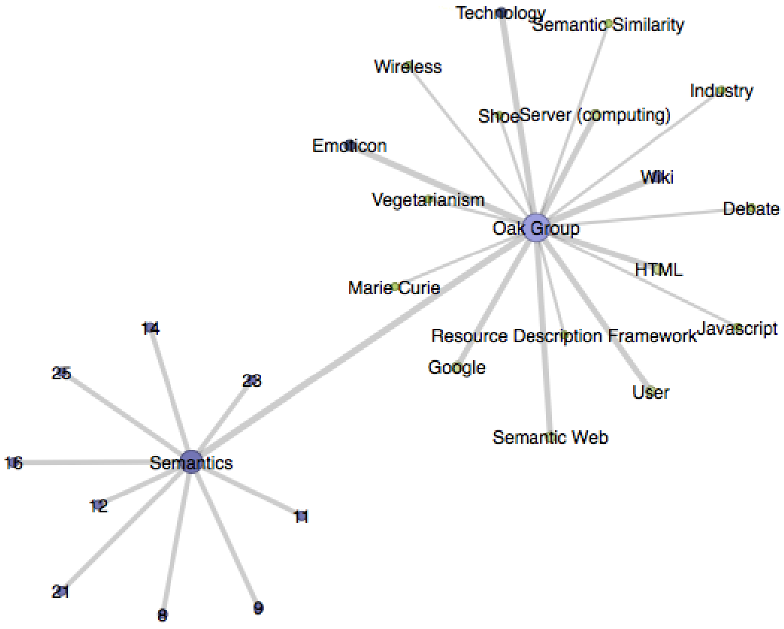

Fig. 2. A force-directed layout to highlight the expertise of the group

Expertise Trend Analysis When wanting to understand the expertise of a group of people or emerging topic trends inside a dynamic organisation it could be helpful to plot all the topics accordingly to their relevance. Figure 2 refers to the 25 users of the experiment shown in section 4. It shows a number of topics discussed within the research group (as extracted from *mlDataset*). The concepts closer to the central node *Oak* are the ones shared by the majority of users, while nodes on the outer circle (randomly picked) are concepts shared by a small number of people (2 or 3). For example, Figure 2 clearly highlights the emerging topics in the group as {*Semantics, Wiki, Emoticon, Technology, User (computing), Week, Semantic Web, Server (computing), HTML, Google*}, but also allows to identify topics that are emerging or have less wide-spread and more specific expertise such as {*Industry, Semantics, Semantic similarity, Javascript, Debate, Vegetarianism, Marie Curie, resource Description Framework, Wireless, Shoe, Chocolate*}.

By using temporal filtering on the data it is also possible to study the trends evolution over time and the hot topics during a certain period. Figure 3 shows the topics discussed in July 2010 and in March 2011 (from *mlDataset*); the data tag cloud visualisation helps discovering that in one period there were discussions e.g. about *Augmented Reality* and about *Facebook* on the other.

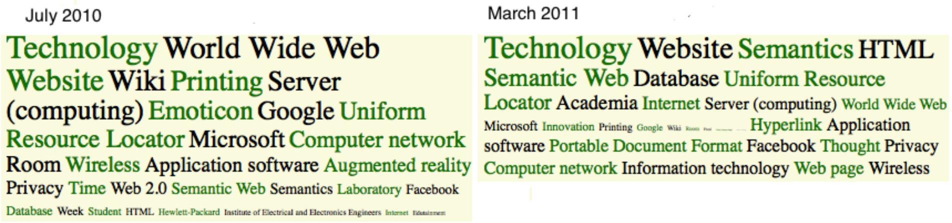
July 2010
Technology World Wide Web
Website Wiki Printing Server
(computing) Emoticon Google Uniform
Resource Locator Microsoft Computer network
Room Wireless Application software Augmented reality
Privacy Time Web 2.0 Semantic Web Semantics Laboratory Facebook
Database Week Student HTML Hewlett-Packard Institute of Electrical and Electronics Engineers Internet
March 2011
Technology Website Semantics HTML
Semantic Web Database Uniform Resource
Locator Academia Internet Server (computing) World Wide Web
Microsoft Innovation Printing Google Wiki Hyperlink Application
software Portable Document Format Facebook Thought Privacy
Computer network Information technology Web page Wireless

Fig. 3. Tag clouds generated over two different period of time in the *mlDataset*.

6 Conclusions and Future Work

This paper proposed an approach to automatically and dynamically model user expertise from informal communication exchanges. The main novelty of our approach consists of generating semantic user profiles from emails (and more generally from any textual user generated content) guaranteeing flexibility, dynamicity and providing ways to connect these data with Linked Open Data (LOD). Whilst linked data have not been exploited in the current work, future work will consider semantic concept expansions, enriching user profiles by exploring the LOD cloud starting from concepts within the profiles. Indeed, the actual concepts in each profile are dbpedia[5] objects, therefore already highly connected to the LOD cloud[6].

Extracting information from informal communication exchanges could be hugely beneficial for knowledge management inside an organisation, as it offers means to recover buried knowledge without any additional effort from the individuals and respecting their natural communication patterns. In order to prove and evaluate the possible benefits for knowledge management the approach has been integrated in a knowledge management system aimed at capturing and sharing knowledge gathered from informal communication exchanges (meetings, chats, etc.) that dynamically builds user networks, determines experts and communities of practice and identifies emerging topics, mirroring the natural evolution of the organisational communities and displays it in an interactive interface that provides means to access and manipulate knowledge. Further evaluations of the approach will be conducted shortly with users to test whether the approach provides advantages over standard knowledge management practices. During this evaluation, attention will be given as whether the approach increases efficiency and efficacy for a given task by presenting in a visual way trends and topics of expertise and providing means to search for experts inside the organisation.

[5] http://dbpedia.org/

[6] http://linkeddata.org/

Acknowledgments

A.L Gentile and V. Lanfranchi are funded by SILOET (Strategic Investment in Low Carbon Engine Technology), a TSB-funded project. S. Mazumdar is funded by Samulet (Strategic Affordable Manufacturing in the UK through Leading Environmental Technologies), a project partially supported by TSB and from the Engineering and Physical Sciences Research Council.

References

1. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing user modeling on twitter for personalized news recommendations. In: Konstan, J., Conejo, R., Marzo, J., Oliver, N. (eds.) User Modeling, Adaption and Personalization, Lecture Notes in Computer Science, vol. 6787, pp. 1–12. Springer (2011)
2. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Semantic enrichment of twitter posts for user profile construction on the social web. In: Antoniou, G., Grobelnik, M., Simperl, E.P.B., Parsia, B., Plexousakis, D., Leenheer, P.D., Pan, J.Z. (eds.) ESWC (2). Lecture Notes in Computer Science, vol. 6644, pp. 375–389. Springer (2011)
3. Adamic, L., Adar, E.: How to search a social network. Social Networks 27(3), 187–203 (2005)
4. Balabanović, M., Shoham, Y.: Fab: content-based, collaborative recommendation. Commun. ACM 40, 66–72 (March 1997)
5. Campbell, C.S., Maglio, P.P., Cozzi, A., Dom, B.: Expertise identification using email communications. In: Proceedings of the twelfth international conference on Information and knowledge management. pp. 528–531. CIKM '03, ACM, New York, NY, USA (2003)
6. Cortes, C., Pregibon, D., Volinsky, C.: Computational methods for dynamic graphs. Journal Of Computational And Graphical Statistics 12, 950–970 (2003)
7. Culotta, A., Bekkerman, R., McCallum, A.: Extracting social networks and contact information from email and the web. In: CEAS 2004: Proc., 1st Conference on Email and Anti-Spam (2004)
8. Daoud, M., Tamine, L., Boughanem, M.: A Personalized Graph-Based Document Ranking Model Using a Semantic User Profile. In: De Bra, P., Kobsa, A., Chin, D. (eds.) User Modeling, Adaptation, and Personalization, Lecture Notes in Computer Science, vol. 6075, chap. 17, pp. 171–182. Springer (2010)
9. De Choudhury, M., Mason, W.A., Hofman, J.M., Watts, D.J.: Inferring relevant social networks from interpersonal communication. In: Proceedings of the 19th international conference on World wide web. pp. 301–310. WWW '10, ACM, New York, NY, USA (2010)
10. Diesner, J., Frantz, T.L., Carley, K.M.: Communication networks from the enron email corpus "it's always about the people. enron is no different". Comput. Math. Organ. Theory 11, 201–228 (October 2005)
11. Eckmann, J., Moses, E., Sergi, D.: Entropy of dialogues creates coherent structures in e-mail traffic. Proceedings of the National Academy of Sciences of the United States of America 101(40), 14333–14337 (2004)
12. Freeman, L.C.: Visualizing Social Networks. JoSS: Journal of Social Structure 1(1) (2000)

13. Gentile, A.L., Basave, A.E.C., Dadzie, A.S., Lanfranchi, V., Ireson, N.: Does Size Matter? When Small is Good Enough. In: Rowe, M., Stankovic, M., Dadzie, A.S., Hardey, M. (eds.) Proceedings, 1st Workshop on Making Sense of Microposts (#MSM2011). pp. 45–56 (May 2011)
14. Gloor, P.A., Laubacher, R., Dynes, S.B.C., Zhao, Y.: Visualization of communication patterns in collaborative innovation networks - analysis of some w3c working groups. In: Proceedings of the twelfth international conference on Information and knowledge management. pp. 56–60. CIKM '03, ACM, New York, NY, USA (2003)
15. Guy, I., Jacovi, M., Perer, A., Ronen, I., Uziel, E.: Same places, same things, same people?: mining user similarity on social media. In: Proceedings of the 2010 ACM conference on Computer supported cooperative work. pp. 41–50. CSCW '10, ACM, New York, NY, USA (2010)
16. Heer, J., Boyd, D.: Vizster: Visualizing online social networks. In: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization. pp. 5–. IEEE Computer Society, Washington, DC, USA (2005)
17. Keila, P.S., Skillicorn, D.B.: Structure in the Enron email dataset. Computational & Mathematical Organization Theory 11, 183–199 (2005)
18. King, D.W., Casto, J., Jones, H.: Communication by Engineers: A Literature Review of Engineers' Information Needs, Seeking Processes, and Use. Washington: Council on Library Resources (1994)
19. Kossinets, G., Watts, D.J.: Empirical analysis of an evolving social network. Science 311(5757), 88–90 (2006)
20. Kramar, T.: Towards contextual search: Social networks, short contexts and multiple personas. In: Konstan, J., Conejo, R., Marzo, J., Oliver, N. (eds.) User Modeling, Adaption and Personalization, Lecture Notes in Computer Science, vol. 6787, pp. 434–437. Springer (2011)
21. Laclavik, M., Dlugolinsky, S., Seleng, M., Kvassay, M., Gatial, E., Balogh, Z., Hluchy, L.: Email analysis and information extraction for enterprise benefit. Computing and Informatics, Special Issue on Business Collaboration Support for micro, small, and medium-sized Enterprises 30(1), 57–87 (2011)
22. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.y.: Mining User Similarity Based on Location History. Architecture (c) (2008)
23. Lin, C.Y., Cao, N., Liu, S.X., Papadimitriou, S., Sun, J., Yan, X.: SmallBlue: Social Network Analysis for Expertise Search and Collective Intelligence. In: Data Engineering, 2009. ICDE '09. IEEE 25th International Conference on. pp. 1483–1486 (2009)
24. Lin, C.Y., Ehrlich, K., Griffiths-Fisher, V., Desforges, C.: Smallblue: People mining for expertise search. IEEE Multimedia 15, 78–84 (2008)
25. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and role discovery in social networks with experiments on Enron and academic email. Journal of Artificial Intelligence Research 30, 249–272 (2007)
26. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceeding of the 17th ACM conference on Information and knowledge management. pp. 509–518. CIKM '08, ACM, New York, NY, USA (2008)
27. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 251–260. WSDM '10, ACM (2010)
28. Pazzani, M., Billsus, D.: Learning and revising user profiles: The identification of interesting web sites. Machine Learning 27, 313–331 (June 1997)

29. Reingold, E.M., Tilford, J.S.: Tidier drawings of trees. IEEE Transactions on Software Engineering 7, 223–228 (March 1981)
30. Robinson, M.A.: Erratum: Correction to robinson, m.a. (2010). an empirical analysis of engineers' information behaviors. journal of the american society for information science and technology, 61(4), 640658. J. Am. Soc. Inf. Sci. Technol. 61, 1947–1947 (September 2010)
31. Schwartz, M.F., Wood, D.C.M.: Discovering shared interests using graph analysis. Communications of the ACM 36(8), 78–89 (1993)
32. Tuulos, V.H., Perkiö, J., Tirri, H.: Multi-faceted information retrieval system for large scale email archives. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 683–683. SIGIR '05, ACM, New York, NY, USA (2005)
33. Tyler, J., Wilkinson, D., Huberman, B.: E-Mail as spectroscopy: Automated discovery of community structure within organizations. The Information Society 21(2), 143–153 (2005)
34. Viégas, F.B., Donath, J.: Social network visualization: can we go beyond the graph. In: Workshop on Social Networks for Design and Analysis: Using Network Information in CSCW (2004. pp. 6–10 (2004)
35. Xiang, R., Lafayette, W., Lafayette, W.: Modeling Relationship Strength in Online Social Networks. North pp. 981–990 (2010)
36. Yee, K.P., Fisher, D., Dhamija, R., Hearst, M.: Animated exploration of dynamic graphs with radial layout. In: Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01). pp. 43–. IEEE Computer Society, Washington, DC, USA (2001)
37. Ying, J.J.c., Lu, E.H.c., Lee, W.c., Tseng, V.S.: Mining User Similarity from Semantic Trajectories. Cell pp. 19–26 (2010)
38. Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.: A comparative evaluation of term recognition algorithms. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Tapias, D. (eds.) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (may 2008)
39. Zhou, Y., Fleischmann, K.R., Wallace, W.A.: Automatic text analysis of values in the Enron email dataset: Clustering a social network using the value patterns of actors. In: HICSS 2010: Proc., 43rd Annual Hawaii International Conference on System Sciences. pp. 1–10 (2010)