Proceedings of the 2nd Workshop on

Semantic Network Analysis

Collocated with the

3rd European Semantic Web Conference

Budva, Montenegro, June 12, 2006

Preface

Recently, collaborative work and social networks came to the fore in many computer science areas. This shows up in the rise of research topics like communities of practice, knowledge management, web communities, and peer to peer. In particular the notion of communities - and thus the need of their systematic analysis - becomes more and more important.

On the other hand, techniques for analysing such structures have a long tradition within sociology. While in the beginning, researchers in that area had to spent huge efforts in collecting data, they nowadays harvest the data free from the WWW. Popular examples are citation and co-author graphs, friend of a friend etc.

A new kind of user-centered applications such as blogs, folksonomies, and wikis, now known as "Web 2.0", consist of large networks of individual contributions, providing a testbed for Social Network Analysis (SNA) techniques at the intersection of the semantic web and SNA areas. The semantic web provides an additional aspect to SNA on the Web as it distinguishes between different kinds of relations, allowing for more complex analysis schemes.

The aim of this workshop is to bring together the semantic web community, the SNA community, and the Web 2.0 community, in order to increase collaboration and exchange of experiences. We assume that the semantic web community can largely benefit from the long tradition present in SNA, and will uncover new possibilities and test beds for semantic technology within the Web 2.0 community.

Besides analysing social networks and cooperative structures within the (semantic) web, our second aim is to exploit the results for supporting and improving communities in their interaction. An important research topic is thus how to include network analysis tools in working environments such as knowledge management systems, peer to peer systems or knowledge portals.

The workshop follows the successful first SNA workshop held at ISWC 2005, and continues a series of workshops on Semantic Web Mining which have been held at ECML/PKDD data mining conferences in 2000-2003 and on Ontologies in P2P communities at ESWC 2005.

We wish to express our appreciation to all the authors of submitted papers and to the members of the program committee for making this workshop an exciting event.

June 2006

Harith Alani
Bettina Hoser
Christoph Schmitz
Gerd Stumme

Program Chairs

Harith Alani (University of Southampton, UK)

Bettina Hoser (University of Karlsruhe)

Christoph Schmitz (University of Kassel)

Gerd Stumme (University of Kassel)

Program Committee

Vladimir Batagelj (University of Ljublijana)

Ulrik Brandes (University of Konstanz, Algorithmics Group)

John Davies (BT)

Tom Heath (Open University, Knowledge Media Institute)

Andreas Hotho (University of Kassel)

Nick Kings (BT)

Sebastian R. Kruk (DERI, Galway)

Enrico Motta (Open University, Knowledge Media Institute)

Kieron O'Hara (U. of Southampton, Intelligence, Agents, Multimedia Group)

Alex (Sandy) Pentland (MIT Media Lab, Human Dynamics Group)

Nigel Shadbolt (U. of Southampton, Intelligence, Agents, Multimedia Group)

Steffen Staab (U. of Koblenz, Information Systems and Semantic Web Group)

Table of Contents

Topic Communities in Peer to Peer Networks
Invited Talk

Steffen Staab

Research Group Information Systems and Semantic Web, Faculty of Computer Science,
University of Koblenz-Landau, Universitätsstraße 1, 56070 Koblenz, Germany
http://www.uni-koblenz.de/ staab/

Abstract. Social communities emerge by successful communication (or other forms of successful reciproke interaction). We have investigated algorithms that forge topic-specific links of different strengths within a peer-to-peer network based on the success of previous communications with regard to the specific or semantically similar topics. We have shown that through such adaptation overlapping topic communities arise that allow for successful routing of queries through peer-to-peer networks.

From semantic to social: an integrated approach for content and usage analysis

Vincent Dubois[1] and Cécile Bothorel[2]

France Telecom R&D, Laboratoire TECH/EASY, Av. Pierre Marzin,
F-22300 Lannion, France

Abstract. Current technologies usually consider documents and their content on one side (Information Retrieval, Internet Search etc.) and their relations with people on the other side (Social Network Analysis, Usage Mining). We propose an automated process integrating both sides in a unified manner in order to exploit heterogenous data available in many enterprises.

We describe an experiment with a project team where we exploit the Intranet files related to the given project and the files usage by project members. A specific aspect of our work is that documents and humans - the employees, the authors - are considered as searchable items. The application we present here is a web site where a user can search for contacts as well as technical data, and explore in a social manner a team project where people and documents are described semantically.

1 Context

Enterprise information is managed and capitalised in data warehouse solutions, where heterogeneous data are stored, integrated, dated, and where they can be accessed within a decision process. The managing of such frameworks is a difficult task and editors are deploying many efforts to achieve it. They center their problematics on the management of documents - from simple text to multimedia ones. Our reflexions lead us to enlight the fact that the human - the employees, the authors - are not taken into account in the enterprise memory process. When we browse our Intranet, it is a challenge to find who can help on a technical problem, who is the guru of a scientific domain, who can lead me to the manager of a project, etc. Given a community working on a project, how to find people and documents related to a technical point? We are interested in solutions unveiling relationships between people, documents and topics. Our challenge is to humanize Intranets, from the Intranet user point of view and include the actors into the data. It is important to store, manage and search the documents, but also to keep a track on who wrote them, who are interested by them. This paper presents a fully unsupervised process combining semantic and social data, and shows how the mathematical object "graph" (or network) is used to integrate heterogeneous data and generate exploitable browsing information.

2 Introduction

The problem is to conceive a methodology and an integrated process which:

- is unsupervized and can be launched every night to maintain fresh information,
- can manipulate a geat volume of data,
- starts with rough data: documents, logs indicating users usage (who fetched which documents),
- integrates heterogeneous data combining both document and human dimensions,
- generates exploitable data representation to allow direct Intranet access: search, browse, personalized information retrieval, centered on documents and/or actors.

A process taking as entries rough data such as users logs and producing ready-to-use information involves different techniques at each stage of treatment. Information retrieval methods can be used to prepare and search documents, but are not relevant for people browsing. Social network analysis is dedicated to social and community network data, but are not efficient on complex and huge networks. Usage analysis and logs (pre-)treatmant can extract similarities between users and documents, but is pointless without relevance and quality information.

In order to answer the challenge of navigation through relations between people, documents and topics, we need to use all these techniques. As a methodology, we study the opportunity to build relation between people and documents and topics early in our process and apply network analysis to cluster, generate relevant relations, remove unrelevant ones, calculate the nodes the more important, etc.

As we propose here to build a unique integrated network using information from document analysis and users usage, we are not presenting special techniques improving information retrieval, nor social network analysis, nor usage mining. Our originality lies in our process: the chaining of existing techniques and the use of "graph" structure as a tool to manage relations between resources (either documents, topics, and people).

3 Related works

Some interesting works, such as WebRefferal [1], acknowledge the usefullness of social network data in information retrieval tasks. When searching for an expert (i.e., an information source), closenes (in the social network) may help in getting searched information. In WebRefferal, the social network is discovered automatically using name cooccurences in public web pages. This technique allows to build network without using usage data (although some usage information is inherently present in the web through co-autorship), but may be prone to some

data discrepancy due to ambigous names and other noises. In our case, usage logs are available and hence we rely preferably on this data source.

The use of logs is well described by Web Usage Mining techniques. They apply data mining techniques to the usage of Web resources, as recorded in Web server logs, helping in traffic reports. Most of the works manipulate the strict URLs and propose typical paths on the Web, URLs frequently associated during the user sessions, shopping histories or navigation histories. As the results interpretation is not evident, studies propose to introduce a preprocessing step to enrich logs with added knowledge. [2] associates user descriptions such as age, gender, professional activity to characterize the resulting navigation patterns. But many works are interested in describing the interest of users and thus need to introduce semantics in the web usage mining process. [3] describes how to enrich logs with semantic knowledge, provided by metatags and RDF annotations and how URLs are mapped into an ontology. Since our study does not use any ontology, we just consider logs to associate contents to users in order to produce profiles coherent with their activities and thus base our community discovering process on practical user description.

When logs are generated by some specific application, such as those designed using the WebML conceptual model, it is possible to take advantage of these *conceptual logs* using some association mining tools, as done [4]. This approach combines analysis of data presents in the logs with informations about the application.

As traditional Semantic Web top-down approaches and dynamic bottom-up ontologies seems irreconcilable, [5] shows how to connect ideas from Social Software, and especially social tagging with the Semantic Web. The former studies and KR&R well fit complex domains such as medicine within expert application based on stable knowledge, but with the Web 2.0 and the Social Software, web applications manipulate unstable and much more dynamic knowledge. P. Mika extends the traditional bipartite model of ontologies with the social dimension, manipulating folksonomy (from folk and taxonomy) such as del.icio.us or BibSonomy, and proposes a model combining actors with concepts and instances. The graph-based process shows how community-based semantics emerges and opens perspectives to propose Semantic Social internet services dedicated to communities with their inherent dynamic ontologies. This work is very close to ours from a methodological point of view. The difference is in its application to Web 2.0 services and the use of folksonomy as tag resources, which could be an interesting extension for us.

4 Data

We describe here an one-year experiment involving one of the FTR&D projects. This project uses a Webdav-based collaborative framework. The following rough data were available:

– a set of 900 technical project-related documents, including reports, user manuals, meeting summary and so on (English and French language)

– 627650 lines of document access logs by 50 people: Webdav format [6] with user identification, time, document, protocol (get,put).

The project is an important technical project on VoIP: its size justifies the need of a dedicated searching tool, allowing to search and browse people and documents related to a very specific topic such as "gateways" or "authentication" for example.

5 Objectives

Using unprocessed data, we want to extract the following information:

– document topics,
– users areas of interest (may be seen as "competencies" in a R&D environment),
– similarities between users and between documents

Moreover, we also want to keep track of some initial data, such as document usage, e.g. relations between users and documents.

As an autonomous process directly exploitable by the end-user, we store the produced information in a relational database using a few tables. The database is not described here, but is accessible by the web site delivering the end-user service [7].

6 Method

In order to compute the relations between resources (documents, topics, and people), we proceed step by step, computing information and then merging the result in a graph as a generic structure. We chose to store the result in progress in a graph structure instead of the database for the following reasons:

– graph structure is intuitive for relationship data,
– graph structure is convenient for most of the algorithms we used: clustering based on neighbouring calculation, semantic propagation,
– a single graph with attributes on edges and vertices can easily represent data that would require many tables,
– a graph with attributes on edges and vertices can easily store heterogeneous data (nodes describing topics, people or documents); the relevant information for us are the relationships: edges may describe usage relation or semantic relation, and filtering tasks can focus on one peculiar type of link.

We describe now the main stages of the process.

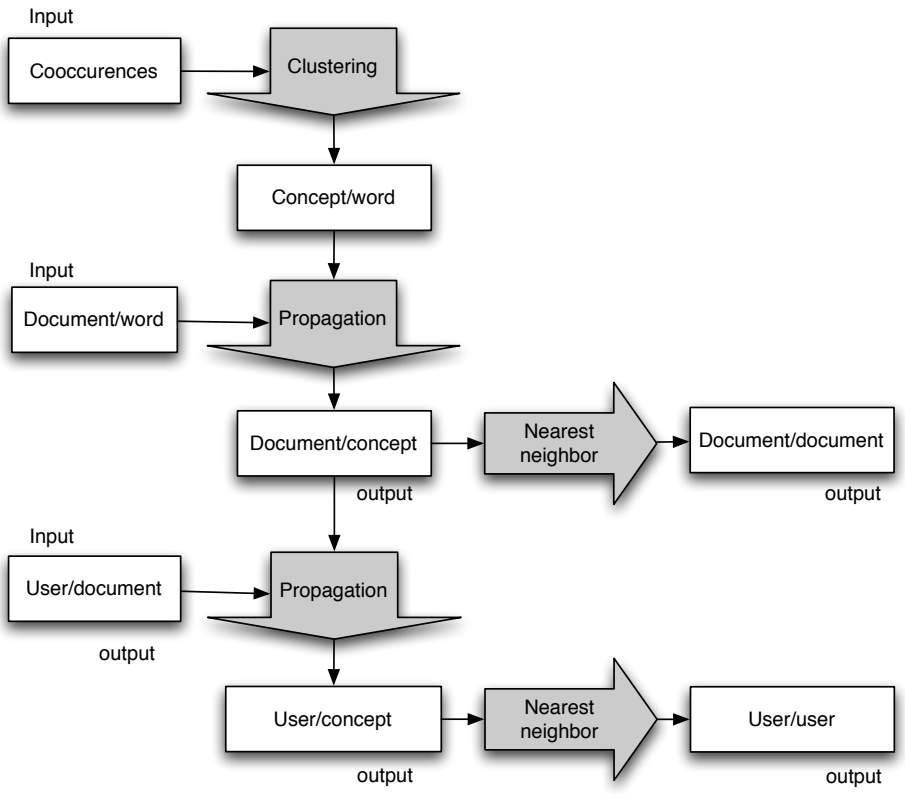


Fig. 1. Global process scheme.

6.1 Global process generating a network of semantically tagged resources

We will now present the global process as shown by figure 1. Each step will be detailed in subsequent sections.

- File analysis: keyword extraction, semantic tagging of documents
- Semantic propagation: associate a semantic description to users according to their document use,
- Profile similarity: compare users and files according to keywords

These three main stages are designed to produce a unique and unified network where users and documents are interconnected. The created relationships are essentially semantic links: connected resources must to be interpreted as "similar" resources according to their semantic tags; especially between users, links mean that users are described with similar keywords. No social relationship is taken into account in this experiment (but let us note that adding the social dimension would be interesting to experiment).

6.2 File analysis

Files are processed in the following way: firstly, each document (".doc" , ".pdf" etc.) is wrapped to textual data. The full collection of texts is used to produce a vector space semantic representation (HAL- like) [8]. To this end, we compute collocations between words, using well-chosen window size, textual units and so on [9].

Secondly, words are clustered in a non-crisp way, so that each word may belong to one or more clusters. Each cluster of words is then used as a "concept", and labelled using the most relevant word in the cluster.

Our clustering method is based on two steps. We first compute a k-nearest neighbour graph using word positions in the vector space. Then we compute clusters using a custom graph-based algorithm. Our algorithm provides generic/specific relationships among elements within each cluster, which is used to automatically label clusters. The main idea of this clustering method is to sort vertices (i.e. words) according to their specificity, using the similarity graph. Words which are considered as neighbour by many other words are more likely to be general and representative terms. Clusters are built around this representative terms by following nearest neighbour links, oriented by the generality relation on words.

6.3 Semantic propagation

We use a simple but efficient way to propagate semantics to documents and users.

The main challenge is that a single word may belong to different clusters, i.e. may have different interpretations. If we want to extract all concepts that appear in a text, we have to remove this ambiguity, which is complex task.

So, the main idea here is that we only want the most significant concepts for each document. We may assume that an interesting concept occurs a significant number of time, and that different words related to this concept will be involved. This make our problem much simpler: we only need to discover which concepts have the most word that may be related. We can achieve that by counting one hit for each clusters a word belongs to. Of course, only one of these clusters deserves the credit, and some concepts will receive undue attention. But, as we are only interested in the most frequent concepts, we ignore concept with low count. On the opposite side, a concept that achieves a high score has many related words in the text, and is very likely to be relevant.

The propagation method is quite simple:

- for each word, we credit each relevant concept,
- we select the most frequent concepts.

In order to keep results balanced, a few improvements have been used:

- for each word getting the same total contribution to the score, we divide the points given to clusters by the number of clusters,
- in order to take differences between clusters into account, we divide each point scored by a cluster by its size; this avoids overweighting large clusters,

– score ceiling is done according to the highest score achieved on each document: for example, any concept with a scoring 0.9 times the greatest score is considered relevant.

This method is used twice in our process: using document/word and word/concept counts, we compute the set of the most important concepts associated with each document. But also, we attribute concepts to users, exploiting the tagged-documents result, we propagate this information to users using user/documents logs. We get then users "interest" or "competencies" profiles with concepts.

6.4 Users and files similarity

In order to present readable and relevant information to end-users, we have to calculate, for a given document, which other documents are similar. In order to extract this information, we use the concepts previously extracted for each document. We consider that document are similar if their distribution of concepts are similar. This can be efficently computed if the measure of similarity is the cosine distance and using our graph structure.

More formally, if we have two documents represented by vectors over concepts (i.e. a value - possibly 0 - is associated with each concept), their distance can be measured by:

$$d(u,v) = 1 - cos(u,v) = 1 - \frac{u.v}{|u|.|v|} = 1 - \frac{u}{|u|}.\frac{v}{|v|} \tag{1}$$

This means that if vectors are normalized, the distance can be computed easily using scalar product. The following algorithm computes the k-nearest neighbours using a bipartite weighted document $\rightarrow$ concept graph.

input : $G(V,E)$; $f : E \mapsto R$
output : knn : $V \mapsto V^k$
for $u \in V$ **do**
  LET score : $V \mapsto R$
  for $(u,v) \in E$ **do**
    for $(w,v) \in E$ **do**
      score$(w) \longleftarrow score(w) + f(u,v) \times f(w,v)$
    end for
  end for
  knn$(u) \longleftarrow k$ vertices with highest score value.
end for

The most outer `"for"` instruction iterates $|V|$ times, the second `"for"` iterates $|E|$ times (for each vertice, the outgoing edge is visited once). So the most inner instruction is executed at most $|E| \times d_{\max}$ times. If score is stored in a hashmap, then the global complexity is the number of edges times the maximum degree. In our application, the degree is the number of important concepts related to a document, and can easily be bounded. This complexity is obviously better than the straighforward computation of every possible distance between vertices ($|V|^2.d_{\max}$).

This result is possible because we take advantage of the sparness of the data by using a hollow structure representation (a graph in our case). In general, computing the k-nearest neighbours on sparse high-dimensional data is a hard problem, and spacial structures such as KD-trees do not help much because of the low number of points with respect to the dimension (they are equals) [10], [11].

6.5 Social browsing within a semantic-tagged network of resources

Information extracted from the data (our unified network) is available using an Intranet application. Figure 2 is a snapshot of the user interface. Our application is developped with the Open Source ePortfolio Framework Elgg [12]

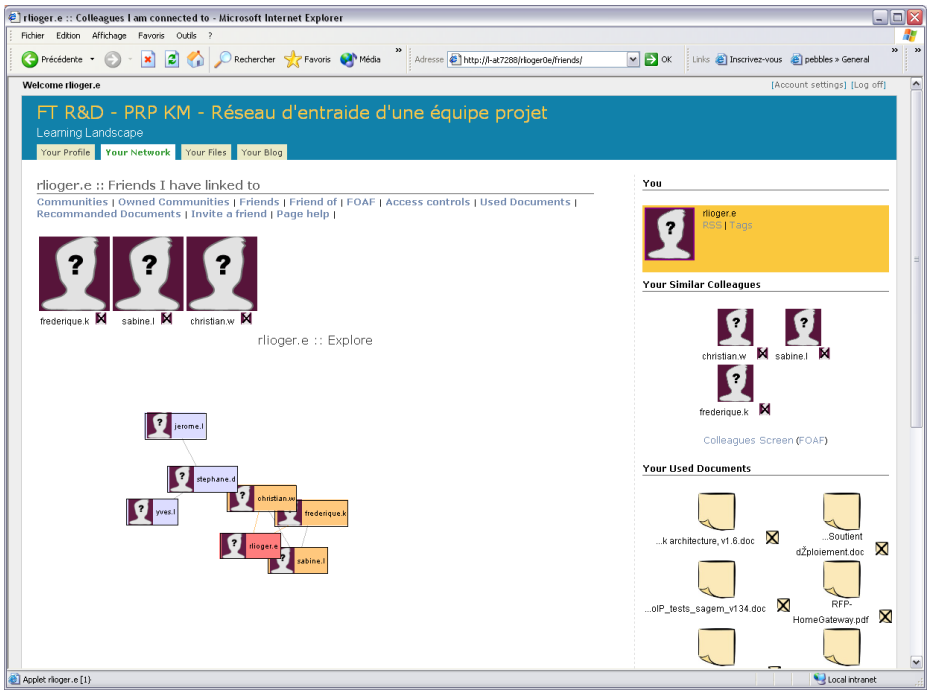


Fig. 2. Intranet portal screenshot

This application is still under development, but it is already possible to search for a document or a contact related to a tag. The results are presented in a network, and thus users are invited to explore in a friendly way their project environment by browsing through their colleagues and documents. We introduce social media functionalities such as visualising global and local community relationships. Digital communities need to have a representation of their global

social organisation but also must support proximate social relations (the Global Village metaphor [13]). Our tool presents the "Global Project Community" as a set of documents but also of people, and shows how they are all interconnected. But the tool provides local views of the whole team/project centered on the current userv (figure 3). We try to reproduce a natural phenomena which encourages people asking information to their close relations even if they are not the experts instead of contacting gurus. Our tool is designed to present the most relevant results in the context of one specific user's relationships, i.e. filtering topic-related documents to the neighbourhood of this user. Neighbourhood reveals here a "semantic" proximity, but we assume that people who works in a similar topic area may know each others. We don't assume any "affinity" proximity but "social" one, meaning here that colleagues working in similar domains are in the same or close or known enterprise environment.

We think that visualising local "semantic" context (e.g. people and documents related to "gateways" topic) help people in situating their professional activity and encourage them in their active participation; but through the network structure, they are invited to apprehend a broader scale representation, and can thus situate their role. We introduce SNA-based scoring methods such as centrality measures (eigenvector centrality [14], [15]) and we think that viewing "importance" of documents or people within the context of a peculiar thematics will favour the Information Retrieval process. By the way, Google use the well-known PageRank algorithm [16] which is a simplified eigenvector centrality to calculate the pertinence of sites. Google proves that topological algorithms are well adapted to score linked resources, and that users are familiar with this way of thinking.

7 Conclusion

We described here a fully unsupervised process dealing with rough data (documents and users usage logs) and filling a database accessible by the end-user web site. Our approach shows how to integrate several well-known techniques (file analysis, clustering, propagation, similarities calculation). We propose also to use a generic graph model to manage and generate relevant relationships between documents, between users, and between users and documents. We suggest that graph structure is efficient to combine heterogeneous data, but also to use techniques based on neighbourhood and distances.

We presented a tool providing browsing within a Project Team Community where users can visualise both their proximate environment and global community. We have just began to experiment SNA algorithms such as centrality and hope that scoring importance of documents and people will bring natural guidelines to users and improve their participation. This combined approach can obviously be used in other domains, such as mailing-list analysis or non-enterprise communities.

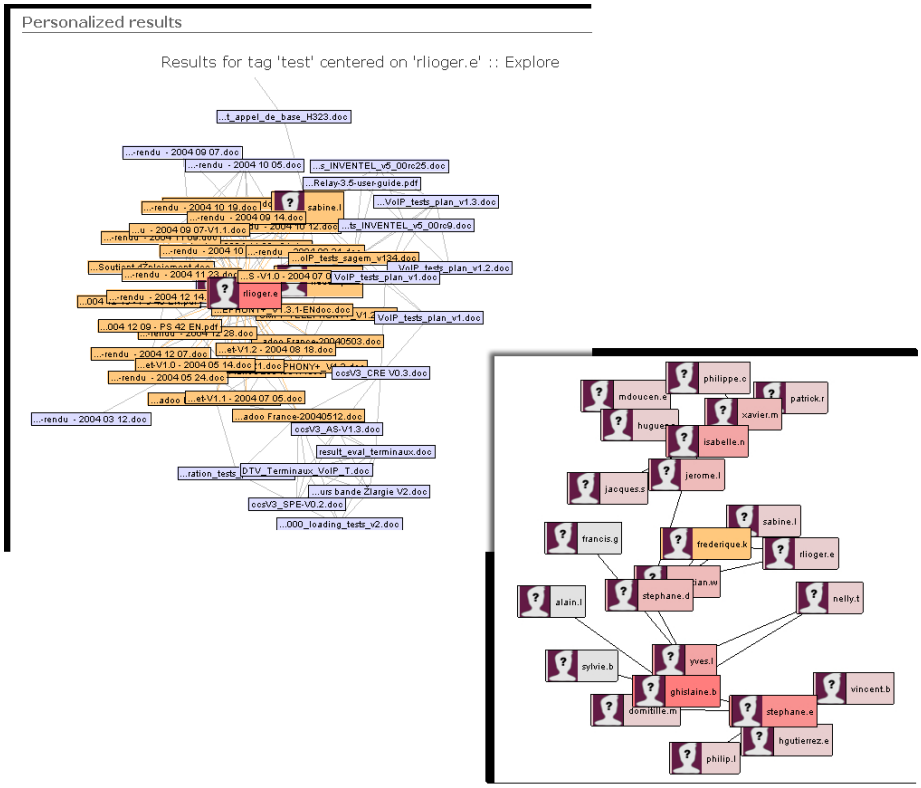


Fig. 3. User centered search: the centered view shows the results to the topic "test" request; the documents shown are directly connected to the user (the orange ones) or in his/her neighbourhood, i.e. connected to connected users. The second screenshot shows the centrality-based scoring where the most central users(documents) are red; centrality is calculated in the local view and has a meaning within the request.

References

1. Kautz, H., Selman, B., Shah, M.: Referral Web: Combining social networks and collaborative filtering. Communications of the ACM **40**(3) (1997) 63–65
2. Chevalier, K., Bothorel, C., Corruble, V.: Discovering rich navigation patterns on a web site. In: Discovery Science. (2003)
3. Oberle, D., Berendt, B., Hotho, A., Gonzalez, J.: Conceptual user tracking. In: First International Atlantic Web Intelligence Conference AWIC. (2003)
4. Meo, R., Lanzi, P.L., Matera, M., Esposito], R.: Integrating web conceptual modeling and web usage mining. In: WebKDD Workshop. (2004)
5. Mika, P.: Ontologies are us: A unified model of social networks and semantics, best paper award. In: 4th International Semantic Web Conference (ISWC). (2005)
6. http://www.webdav.org/: Official site (2006)
7. Bothorel, C., Dubois, V., Van Coillie, M.: Social knowledge management for new entreprise km paradigme. In: ePortfolio. (2005)

8. Lund, K., Burgess, C.: Producing high-dimensionnal semantic space from lexical co-occurrence. Behavior Research Methods, Instruments & computers **2**(28) (1996) 203–208
9. Meyer, F., Dubois, V.: Exploration des paramètres discriminants pour les représentations vectorielles de la sémantique des mots. In: EGC. (2006) 275–286
10. Bentley, J.L.: Multidimensional binary search trees used for associative searching. Communications of the ACM **18** (1975) 509–517
11. Samet, H.: Applications of spatial data structures. Addison-Wesley (1990)
12. http://www.elgg.net/: Official site (2006)
13. Wellman, B.: Networks in the Global Village. Wellman B., Westview Press, Boulder, CO (1999)
14. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology **2** (1972) 113–120
15. Brandes, U., Cornelsen, S.: Visual ranking of link structures. Journal of Graph Algorithms and Applications **7(2)** (2003) 181–201
16. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Journal of Graph Algorithms and Applications **30(1-7)** (1998) 107–117

Representing Social and Cognitive Networks*

Wouter van Atteveldt[1,2], Jan Kleinnijenhuis[2], Dirk Oegema[2], and Stefan Schlobach[1]

[1] Department of Artificial Intelligence, Vrije Universiteit Amsterdam,
{wva,schlobac}@cs.vu.nl
[2] Department of Communication Science, Vrije Universiteit Amsterdam,
{j.kleinnijenhuis,d.oegema}@fsw.vu.nl

Abstract. Graph-like data structures are central to the fields of Social Networks Analysis and Relational Content Analysis, and form the semantic backbone of Semantic Web (SW) and related formal representations. Taking a Social Science perspective, we will make an inventory of the possibilities and problems in using Knowledge Representation (KR) and SW techniques for the modeling of and reasoning with social data. We also present a prototypical implementation of such a modeling. Our conclusion is that the main difficulties are transparently extending triples with added (meta)data, representing quantitative aspects of relationships, and recursive embeddedness of networks. These challenges notwithstanding, we think that the formalization of social science data has great advantages and will become increasingly important as studies become more complex and combine more different data sources and methodologies.

1 Introduction

Graph-like data structures are central to the fields of Social Networks Analysis, Relational Content Analysis and the Semantic Web. Social Networks Analysis (SNA) generally considers the social interactions or communication structure between actors [1]. Each communicating actor transmits texts that lift a corner of the veil of a particular cognitive network at a particular point in time. Relational Content Analysis (RCA) deals with extracting the relationships between actors, issues, values and facts from texts, for example relationships of support/criticism or cooperation/conflict between actors, issue positions as relationships between actors and issues, or subjective causal relationships between issue developments [2–7]. Knowledge Representation, and especially the interrelated fields of Modal Logics, Description Logics, and the Semantic Web, concernt themselves with representing and inferencing over graph-like data such as RDF graphs or Kripke Structures, and formally describing properties of or constraints on this data [8–10].

This resemblance notwithstanding, these fields operate relatively isolated from each other. The desire to integrate the study of communication networks between actors [11] with the study of the cognitive networks of actors was already apparent in 1986 [12], but SNA is still primarily concerned with relations between actors, whereas RCA is mainly targeted at studying the separate cognitive networks of the media and powerful political actors. Finally, the Modal Logics KR community is mainly interested in studying the formal properties of graphs and logical formulas and the relations between them.

This paper will not present clear solutions or scientific findings of applying KR techniques to social data. Rather, it will investigate the possibility of combining results from these fields, specifically by looking at Semantic Web techniques

* We would like to thank the reviewers for their honest and constructive comments

as a solution for interoperability problems in Relational Content Analysis. We will survey which aspects of Relational Content Data can be easily modeled, and which aspects cause more problems. Finally, we will present a first concrete prototype application using those aspects that can be easily implemented.

1.1 Relational Data in Different Disciplines

Representation of SNA and RCA data in the data structures that were developed recently by the KR community, such as RDF, appears to be desirable. In the first place, formal KR techniques allow us to have declarative semantics of our networks, allowing for easier sharing and combining of data. Moreover, the newlydeveloped data structures promise access to a wide range of tools for logical, statistical and graphical analyses. This would be an enormous improvement as compared to the common practice of either using multiple-purpose statistical programs which do not reckon with the precise nature of network data, or using relatively ad hoc stand alone programs to analyze network data and likewise formats to share the data.

In general, representing graphs in RDF is a trivial matter, and the subtype mechanisms in RDFS provide for formal definitions of vocabulary in terms of shared concepts as required for combining heterogeneous data sources. However, two features present in certain social and cognitive networks analyses pose problems: Recursion/embeddedness of networks and nominal or quantitative aspects of complex relationships.

Recursion in this case means the embedding of one network within another, where the nodes of the second network are networks themselves. For example, actors, who are the nodes in a social network, have a cognitive network, which contains a representation of the social network that surrounds them, which recursively consist of nodes which contain perceived cognitive structure of these actors. Although RDF has some support for reification this is generally criticized as a weak feature (eg [13]). Multidimensional logics allow graphs to be embedded in graphs, but this is still very much work in progress. Note that within SNA and RCA embeddedness is still a substantive and methodological challenge, so this can be seen as a joint problem rather than a solution looking for a representation.

The other problematic feature is that relationships are often complex, with different quantitative and qualitative aspects that need to be represented. Relationships may belong to different classes or categories such as willing, acting, and causation. Additionally, pervasive aspects of social relationships in a particular time span, such as the frequency or duration of contacts, the average degree of agreement/disagreement, their ambiguity, or their inconsistency are represented as real numbers. RDF(S) is highly suited for representing different nominal aspects of relations using the subproperty mechanism, and describing quantitative aspects of nodes is possible using the typed literal system. However, representing quantitative aspects of relations is difficult in RDF as relations are not allowed to be the subject of other relations, and no mechanism exists for valued or complex relations.

1.2 Structure of this paper

This paper starts with two recent studies from the field of Relational Content Analysis to indicate the needs to be addressed by any formal Knowledge Representation of SNA and RCA. The first case study, described in the next section, presents an analysis of the reciprocal influences between internet forums and mainstream news media with respect to the immigrant issue in the Netherlands from 2001-2005 [14]. This section outlines the challenges connected with combining data at different levels from different sources.

The subsequent section is based on an analysis of media coverage of the 2002 election campaign in the Netherlands, which was dominated by the immigrant issue and the assassination of Pim Fortuyn [15]. This case study uses a method with complex relations with both quantitative and qualitative aspects.

The last substantive section will briefly describe a way to model this data in an RDF framework. This will make the problems mentioned earlier more concrete and show which aspects can be very successfully modeled, highlighting the benefits for the RCA and SNA community in adopting these methods.

2 Emotions and Influence on the News and the Net

Discussion forums[3] reduce the costs of publicly articulating private feelings, making them a much used vehicle for expressing (minority) views on controversial issues. Discussion forums therefore provide interesting data for media effects studies. In this section we will describe a recent exploratory investigation of emotions and issues on two Dutch internet forums in connection with the issue of immigration.

2.1 Data and Operationaliztion

The research question is how emotions and impassioned expressions from individual cognitive networks interrelate with the social influence relations between the mass media and the discussion forums.

We expect discussion forums to show a selective expression of negative emotions, polarization and flaming. From the Agenda Setting hypothesis, which states that issues receiving much media attention will be perceived as important, it is furthermore expected that the topics of the postings to be strongly influenced by the topics written about in the news [16, 17].

We collected postings of the two on-line discussion groups from October 1st 2003 through to July 31st 2005. Also, newspaper articles containing relevant actors or issues from the five biggest Dutch national newspapers were retrieved from the LexisNexis archive. The measurement units for the automatic content analysis were separate postings and news articles. A random sample of 30% of all Internet postings was drawn (28,671 and 38,212 postings for `marokko.nl` and

[3] We will use discussion forums and Internet forums interchangably to refer both to web-based forum systems and Usenet groups

The immigration issue in the Netherlands
Islamic immigrants became the foremost controversial issue in the political debate in the Netherlands after 9/11. Besides the events in New York, this was largely due to Pim Fortuyn, a charismatic newcomer to the political scene characterized by the New York Times as "a gay right-wing populist who stood out in the gray political world of the Netherlands". During much of the 2002 election campaign he attracted more media attention than any other politician, voicing radical opinions that until then had been regarded as taboo in the liberal press.
Examples of such statements where his promise that "no Muslim will come in", and his statements such as "Islamism is a backward culture", and "I want to get rid of the constitutional prohibition of discrimination". Attention for the immigration issue rose to a climax in May 2002, the election month during which Pim Fortuyn was assassinated by a left-wing animal rights extremist.
'Never speak ill of the dead,' and accordingly the press and the vested parties alike suddenly started praising Fortuyn for having torn down the taboo of seeing immigration as a problem. This stimulated a revaluation of free speech over political correctness, leading to an increasingly polarized debate. The film Submission, in which Ayaan Hirsi Ali and Theo van Gogh projected Quran texts onto the body of a naked woman being tortured, and the subsequent killing of Van Gogh in November 2004, was fuel for further polarization.
Although the tone in the newspapers remained civil, this polarization is strongly reflected (or even overrepresented) on the internet forums, as two excerpts from the internet forums investigated for this research show:

From `nl.politiek`, a right-wing 'white' usenet group:

> *It's all right with me if they shoot to grits the complete Islam and all the *** dogs who belong to it. ***, dead for all of you, filthy headscarf *** and stinking blokes. *** the goats in your own rotten country. AND NOW *** OFF, ALL OF YOU. PHEW!*

From `marokko.nl`, a Dutch language Moroccan discussion forum:

> *Filthy *** Soussian glutton, may your gullet be given an acute *** and your *** make a somersault through your cellulitic *** as far as up your bulging eye. [...] everything has its own limits thus FREEDOM OF SPEECH too*

`nl.politiek` respectively) while for the national newspapers all articles were collected (n=40,429 articles).

In a methodology comparable to [18], disgust, hatred, shame, love, eagerness, pleasure, polarization and flaming were measured using an automated content analysis with a list of all the direct synonyms of these emotions as their indicators. Disambiguation conditions were formulated to ensure that they occurred in an emotional context. To test the agenda setting hypotheses that the mass media influence the discussion forums, we specified a list of actors and issues

Table 1. Operationalization of and attention for a selection of actors and issues

Row attention scores are percentages, (sub)totals are frequencies

Cluster	Example keywords	#cl	#obj	newsp.	nl.politiek	marokko.nl
government	Prime Minister, minister	1	56	27.8	12.8	10.6
rightist ideologues	Ayaan Hirsi, Van Gogh	1	12	20.1	36.6	37.9
judicial power	Supreme Court, judges	1	10	6.6	3.7	6.0
Islamist extremists	Mohammed B., Samir A.	1	14	3.8	3.2	7.9
immigrant pressure gr.	immigrant interest groups	1	15	1.5	.8	.8
...	...					
subtotal actors		**17**	**256**	**157033**	**39006**	**17058**
terror	terror, 9/11, Madrid	2	20	33.4	26.3	18.2
crime	crime, murder, rape	8	105	19.8	17.3	11.5
Islam	Islam, Quran, mosque	1	14	14.0	19.7	40.6
War Iraq, EU enlrgmnt	enlargement EU, Schengen	1	24	5.2	1.4	.7
Christianity	Christianity, churches	1	2	3.6	2.9	2.8
employment	(un)employment	1	3	.3	.2	.1
...	...					
subtotal issues		**25**	**266**	**245723**	**49229**	**54502**
total actors and issues		**42**	**522**	**402757**	**88235**	**71560**

divided into clusters for which we defined keywords and disambiguation conditions. Table 1 lists a subset of the 42 clusters of actors and issues we used and the number of objects per cluster as well as some example keywords, and the frequencies of these clusters in the newspapers and forums.

2.2 Results

The first object of our investigation is the frequency of verbal indicators of emotions, polarization per 10,000 words, presented in Table 2. As was expected the three emotions with a negative action tendency towards others - disgust, shame and hate - are more prominent in discussion forums than in newspapers. Polarization and flaming are more prominent in the discussion forums also. Disgust and hatred are even more prominent on `marokko.nl` than on the right-wing discussion forum `nl.politiek`, but polarization and flaming are most prominent on `nl.politiek`. Participents at at `marokko.nl` express positive emotions less often than the Dutch newspapers or the users of `nl.politiek`.

Table 2. Number of linguistic stylistic markers of various types per 10.000 words

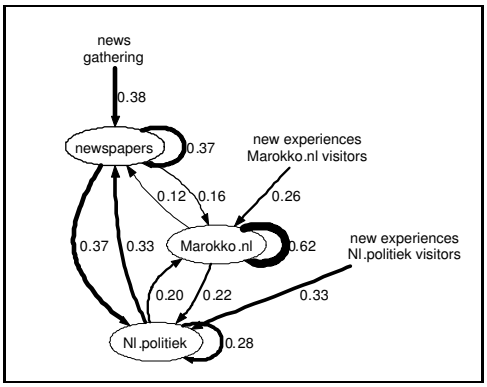
hypotheses	emotions	examples	newsp.	nl.politiek	marokko.nl
positive emotions	love	love, sympathy, tenderness	2.9	3.0	2.5
(control)	eagerness	eager, passionate	0.7	0.6	0.4
	pleasure	pleasure, sensuality	2.5	3.6	2.6
negative emotions	shame	ashamed, embarrassed, faulty	2.5	4.2	4.2
	disgust	distaste, aversion, nauseating	5.5	6.4	7.2
	hatred	hate, absolute disdain, revenge	3.4	5.3	5.8
polarization	polarization	crazy, lunatic, total, enormous	5.6	8.7	5.0
flaming	flaming	racial, religious, bestial,Berber	4.1	21.5	11.6

The next question is: How the mass media and the discussions forums influence each other? For the sake of brevity the threads or interactions between indi-

vidual participants in a discussion will not be discussed here (cf [19]). Correlation coefficients between the attention for actors and issues in the mass media and the discussion forums indicate a high level of interrelatedness ($0.60 < r < 0.77$, n= 522 objects in 22 months=11,484 units of analysis). The same holds for correlations between the mass media and the discussion forums with respect to the attention for specific associations between actors and/or issues ($0.75 < r < 0.60$, n=561 associations between clusters). Although the discussion forums might disagree strongly with each other qualitatively, the agreement with respect to the agenda of relevant issues found here is relatively strong; they might have different opinions but they have opinions on the same subjects. The most important difference between topics in the mass media and the discussions forum is that the latter focus on a narrower set of actors and issues.

Correlations do not prove causality. A structural equation model was estimated to assess the causal order of the agendas of the newspapers and the two discussion groups. The assumption underlying such a model is that each of the three agendas is determined autoregressively by its own recent past (i.e., the agenda of last month) as well as by the current editions of the two other agendas. A visual representation of this model is given in Fig. 1, where the boomerang arrows indicate autoregression and the arrows without source are the unexplained variance.

Fig. 1. A reciprocal model of first order agenda setting: who influences whom?



Each of the estimated parameters is highly significant, whereas the fit of the complete model is acceptable (CFI=0.91, df=3). A comparison of the top-down estimates (from newspapers to the discussion forums) with the bottom-up estimates (from discussion forums to newspapers) shows that top-down agenda setting is not dominant, as was expected on the basis of earlier research on agenda setting. Bottom-up influences are significant, especially for `nl.politiek`. The equations show that `marokko.nl` and `nl.politiek` play an exhaustion game

(0.22 and 0.20), with `nl.politiek` getting more fresh blood from the media every month than `marokko.nl` (0.37 as compared to 0.16), but `marokko.nl` getting far more regeneration from its own past (0.62 as compared 0.28).

2.3 Forum Representation

In representing internet content we automatically measured two separate things per posting: the emotional value and the topics and actors discussed. From this, an association graph was created where the edges are the associations made between topics. For further research, these edges will be enriched with the average emotional content of postings associating the two. Furthermore, a taxonomy was defined over the vocabulary, which allowed us to collapse similar nodes, creating a more parsimonious graph and hopefully reducing noise.

We would expect a representation of this data to store both association strengths and emotional value per posting, and to allow us to query for the average associations and emotions per actor or cluster pair for a given period of time.

In this study, the media content was investigated using the same method as the forum content. However, there are existing corpora of annotated newspaper content. Often, we want to combine new data with existing data, possibly using a different vocabulary and methodology. Formalizing the vocabulary of both studies into an ontology an important first step towards combining their data.

For future research, it would be very interesting to look at the internet forum not as a single association graph but as an interrelated collection of personal cognitive maps. Such a representation potentially allows for investigating the interaction between personal world view, overall forum content, and mass media content. Although such an analysis would pose many substantive problems as well, representing this data in a sensible manner might be a first step toward systematic exploratoin.

3 Cognitive networks presented by the media

This section will show a more complex Relational Content Analysis of the campaign for the 2002 Dutch parliamentary election. Whereas the previous section focussed on frequency and co-occurrence, the method in this section also considers qualitative and real-valued quantitative aspects such as the nature (causation, disagreement) and degree of the relation between actors. Additionally, embeddedness arises from the introduction of cognitive networks according to quoted or paraphrased sources.

3.1 Data and operationalization

Relational Content Analysis uses a graph structure as a text representation, generally having relevant actors or issues as nodes and their relations according to the text as edges. The Relational Content Analysis method used here is the

Political Campaining and the Immigrant Issue
As an example of Relational Content Aanlysis, we will look at the 2002 landslide election victory of the assassinated Pim Fortuyn. Fortuyn succeeded in putting the immigrant issue and its connection to crime, Islam, and terror on the political agenda, taking many votes away from the established political parties who shunned these sensitive issues [15].
Initially, the incumbent coalition parties PvdA (Social Democrats), VVD (Conservatives) and D66 (Liberal Democrats), very much aware of the danger Fortuyn could pose, tried to keep him out of the debate. Instead, they emphasized instead 'serious' differences of opinion between the incumbent parties themselves along the traditional division line between liberal and conservative parties.
In the meantime, Fortuyn was still getting media attention and continued to rise in the polls. On February 9, three months before the elections, he gave an interview airing a number of extremist views, after which the incumbent parties decided that Fortuyn's views were unacceptable in the political debate; his statements allegedly went too far, branding him as a right-wing extremist apparently unaware of what had been done to Anne Frank. This, of course, only served to put Fortuyn's issues firmly on the agenda.

NET method, which is an elaboration of the more generic Evaluative Assertion Analysis developed earlier [4, 20, 3].

In the NET method, each analyzed sentence is represented as a number of nuclear statements, which are essentially triples containing the [$Subject$; $Connection$; and $Object$] of the statement. The Connection in turn consists of a nominal Type, such as associative, causative or affinitive, and a quantified Quality ranging from -1 (total dissociation or dislike) to 1 (maximal association or support). The triple can have an optional quoted source, turning it into a quadruple or embedded triple.

For example, the sentence "The President has been accused of corruption by the house minority leader" is represented by the quadruple [$MinorityLeader$: [$President$; $(ACT + 1)$; $Corruption$]], which shows us a glimpse of the cognitive network of the house minority leader. From this embedded triple we can infer the surface triple [$MinorityLeader$; $(AFF - 1)$; $President$], since corruption is known to be negative. Moreover, depending on the research question, we might want to conclude [$President$; $(ASS + 1)$; $Corruption$] (rules for these inferences are described in [20]).

Although some of the information in the text is inevitably lost in the NET coding process, applications of this method have shown that much relevant information can be systematically extracted and quantified. An example of such information could be the average degree of cooperation or conflict between two specific actors in a given time span. Based on the frequency of interaction and the average degree of cooperation or conflict between two actors, division lines

between actors in the network can be established, for example for partitioning a network into coalitions or 'blocks' [21].

For the period from September 2001 until the elections of May 2002 a relational content analysis was performed, based on the complete items for three televisions news magazines and the headlines and leads for five national newspapers. This resulted in 35,031 $[source : subject; predicate; object]$ quadruples, 12,664 from television news and 22,367 from the headlines and leads of newspapers. The vocabulary consisted of roughly 750 different nodes representing actors (e.g. MPs, ministers, mayors, advisory councils) and issues (e.g. social security, crime). Here the emphasis will be on relations of cooperation and conflict between the political parties, with the focus on the question whether the division lines between political parties shifted after February 9 when the established parties could not ignore Fortuyn anymore.

3.2 Results

The content analysis of newspaper and television news reveals indeed that the media from February 9 onwards replaced the dividing line between leftist and rightist parties that was apparent from the news with a line between Pim Fortuyn on the one hand and the vested coalition parties on the other. This shift resulted in a dramatic loss for the incumbent parties, and an unprecedented election victory out of the blue for Pim Fortuyn's party, and for the CDA (Christian Democrats), who were the only other party to address the immigrant issue, albeit in softer tones ('the multicultural society is not something to strive for'). A graphical representation of the conflict network between the major parties is given in Figs. 2a and 2b.

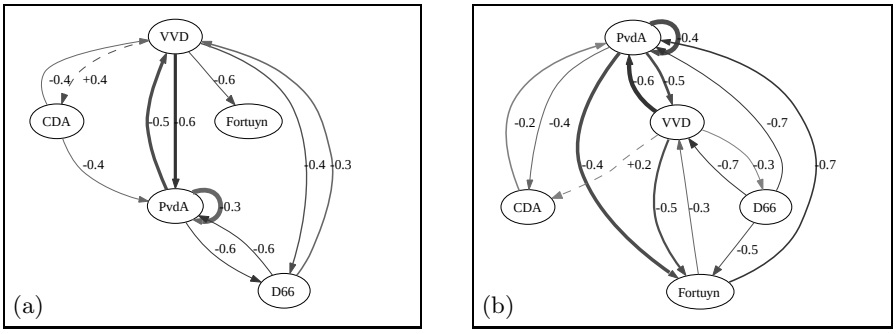


Fig. 2. Relations between the major parties before (a) and after (b) the February 9 interview; dashed lines indicate a positive relationship, grayscale indicates intensity and width indicates frequency.

3.3 Representing Relational Content

Using a manual Relational Content Analysis technique, we were able to construct a very detailed picture of the relations between different actors and nodes during a campaign. This results in a manually created graph per article of complex relations (consiting of a nominal type and real-valued strength) between actors and issues. Complicating this further, quoted statements in an article create separate embedded graphs for the quoted source, with recoding rules interfacing between the two graph layers.

What we would expect of a representation language is a way to represent both the nominal and real-valued aspects of these relations. Moreover, we would want to be able to query this data to summarize these results for a selection of nodes and articles (such as all articles per week per newspaper). Finally, we want to be able to know to which node in an article graph the 'quoted source' (sub)graphs are connected, and ideally express the recoding rules as inference rules.

Moreover, analogous to the previous section, if we can formally represent the vocabulary as an ontology and link it to our background knowledge (eg the knowledge that Wim Kok was prime minister from 1994-2002) this will make it easier to combine different data sources and define research questions formally, at a high-level, and independent of the specific vocabulary used for the study.

4 Towards Knowledge Representation of SNA and RCA data

The basic data model for SNA and RCA is the graph. Hence, it makes sense to look for a representation language that uses a graph as the basic representation, even when the task is to deal with embeddedness and with a variety of (both categorical and real-valued) aspects of relations.

An example of such a language would be modal logic, the semantics of which are based on Kripke Structures. In fact, earlier work finds that Hybrid Logics, an extension of Modal Logics allowing for limited variable binding, allows for the extraction of interesting patterns from graph representations of newspaper articles [22].

However, Modal Logics are inherently ill-suited for representing and manipulating quantitative data, such as the Quality aspect of NET connections, e.g. the degree of cooperation or conflict between actors. This also makes Modal Logics an impractical tool for extracting aggregate or quantitative data such as simple questions with regard to the average degree of cooperation or conflict as posed in the previous section. Moreover, the need for the variable binding of hybrid logics to represent cyclic patterns leads us into a language family with little tool support.

To counter these drawbacks, RDF(S) can be used as a representation language for relational news content [23]. This representation, which is a W3C standard with a broad user base and tool support, also represents information

as a labeled graph, where both edges and nodes can be typed using a RDF(S) hierarchy [8]. Here we will sketch the outline of this approach.

Translating the core structure of NET into RDF is straightforward due to the highly similar data models: a statement is encoded as an RDF statement from the Subject to the Object. Here, the connection is simplified by joining a dichotomized quality and the type as a property, resulting in connections such as 'NegativeCausation' for the original pair ('$Causative'$, -1)

This immediately allows us to link the media data to an RDFS ontology containing the background domain knowledge, featuring information such as 'a minister is a politician'. Among other things, this allows for easy and well-founded aggregation of data from the specific textual measurements to more high-level (and less task-specific) theoretical concepts. This by itself can be a very imporant advantage to the SNA and RCA community.

To track changes to the network or to be able to attribute statements to particular media we also need to encode meta-data about these statements. Although this can be done fairly easy using the RDF reification mechanism, it has two serious drawbacks. The first is that the RDF specification is intentionally vague on the semantics of reified statements, causing the community and tools to mainly neglect this mechanism. Second, reified statements do not imply the original statement, so coding metadata in this way means that non-reification-aware tools or users will be unable to access the abovementioned 'core' part of the data for which reification is not required. Effectively, the declarative semantics of our core network are lost by using reification. Unfortunately, the only other obvious solution within the original RDF framework is using n-tuples rather than triples as suggested by [24], which also disturbs the original data by inserting dummy nodes.

Finally, a mechanism is required to represent the quantitative nature of the relationship. Effectively, we want to add an attribute to the relation rather than to one of the nodes. This places us in the same position as above: both using reification and using n-tuples via dummy nodes means losing declarative semantics and (therefore) generic tool support.

Given these considerations we decided to use reification for both the meta-data and the quantitative aspects (see also [25]). Although far from being an optimal solution, this at least keeps us within the RDF specifications. Moreover, adding extra data to existing triples seems to be a bona fide use of reification. Furthermore, this is structurally equivalent to one of the dummy-node solutions proposed by Noy and Rector, making the `rdf:statement`the dummy node. Finally, if at some point RDF is extended with a context mechanism, the data model can be translated straightforwardly. The RDF schema using reification is visualized in Fig. 3a.

Using this schema to translate the data of a recent study of technological risk coverage regarding the placement of UMTS broadband transmitters, we were able to use SeRQL to define and extract high-level domain independent patterns over this relational data. Moreover, using a prototypical aggregation engine for SeRQL queries allowed us some basic manipulation of the quantitative

data. A screenshot showing how a specific pattern is detected in an article using prototype is shown in Fig. 3b; please see [23] for more details.

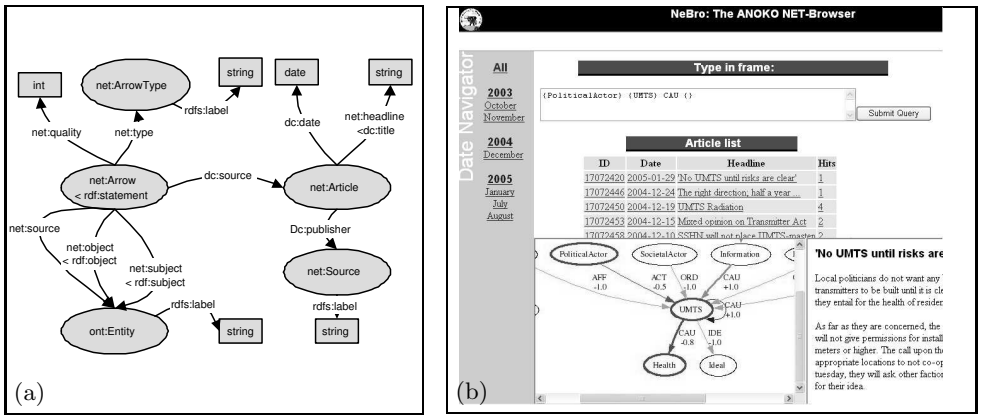


Fig. 3. (a) The RDF schema used; (b) a screenshot of the system showing a pattern detected in an article using an aggregate SeRQL query

However, the goals of sharing our data across communities and linking it to existing tools and data sources are strongly compromised by the use of a semantically difficult mechanism (reification) for two conceptually different goals, namely quantifying relations and representing metadata [23]. Recent proposals in the RDF community for adding a Named Graph context mechanism to RDF could solve some of these problems, although the clarity of using contexts for coding both quantitative aspects and metatada is questionable [13].

This implementation intentionally ignored embedded networks such as quoted sources since reifying a reified statement is impossible. However, as Named Graphs as proposed by [13] do allow for nested graphs, their proposal might be more promising in this respect.

5 Discussion

Since Social Networks Analysis, Relational Content Analysis, and Semantic Web techniques all use graphs as a modelling primitive, we expect that much can be gained by combining results from these fields. This paper summarized two recent Relational Content Analysis studies, one using automatic text analysis and one using manual annotation, and a proposed method for formalizing the structures from these analyses in RDF.

This formalization was hampered by two features of the investigated content analysis methods: Quantitative aspects of the relations and embedded graphs. Nonetheless, a model was created using the reification mechanism that allowed

us to express and query the data in a satisfactory manner, although it ignored the embedded aspects. If a mechanism is added to RDF that allows for nested Named Graphs, such as proposed in the Semantic Web literature [13], it will be easy to transform the presented data model into a more elegant model using these Named Graphs, taking away some of our objections. Moreover, this would also make it possible to represent the embedded graphs.

One of the biggest advantages of settling on a formal data model that allows for a good definition of the used vocabulary, such as RDFS, is that this will greatly increase the ease with which data from different sources can be exchanged and combined. As we expect the necessity for studies combining data to grow as a consequence of the increasing sophistication of Communication Science theory and the increasing interconnectedness of communication, this formalization will become ever more important.

In conclusion, we would like to state that we think there is a great future for the use of formal KR methods in the Social Sciences, and in particular in the Relational Content Analysis and Social Networks Analysis, both due to the relative complexity of the data structures involved and the close match of these data structures to the graph structures used in Semantic Web (and related) formalisms. However, more work is certainly needed in terms of support for real-valued data, and for enriching existing triples in a transparant way. Finally, recursive use of networks as nodes in other networks, which is a daunting challenge in both substantive theory and from the perspective of Representation and Reasoning, is a field where much might be gained from collaboration between these fields.

References

1. Wasserman, S., Faust, K.: Social Network Analysis. CUP, Cambridge (1994)
2. Kleinnijenhuis, J.: Applications of graph theory to cognitive communication research. In Krippendorff, K., Bock, M., eds.: The Content Analysis Reader (forthcoming). Sage, Thousand Oaks (2006)
3. Kleinnijenhuis, J., De Ridder, J., Rietberg, E.: Reasoning in economic discourse: an application of the network approach to the Dutch press. In Roberts, C., ed.: Text Analysis for the Social Sciences; Methods for Drawing Statistical Inferences from Texts and Transcripts. Lawrence Erlbaum Associate, Mahwah, New Jersey (1997) 191–207
4. Osgood, C., Saporta, S., Nunnally, J.: Evaluative assertion analysis. Litera **3** (1956) 47–102
5. Popping, R.: Computer-assisted Text Analysis. Sage, Newbury Park / London (2000)
6. Schrodt, P.: Automated coding of international event data using sparse parsing techniques. In: Annual meeting of the International Studies Association, Chicago. (2001)
7. Young, M.D.: Building worldviews with profiler+. In West, M.D., ed.: Applications of Computer Content Analysis. Volume 17 of Progress in Communication Sciences. Ablex Publishing (2001)

8. Antoniou, G., van Harmelen, F.: A Semantic Web Primer. MIT Press, Cambridge, Ma. (2004)
9. Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Volume 53 of Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge (2001)
10. Daconta, M.C., Obrst, L.J., Smith, K.T.: The semantic web. Wiley, New York (2003)
11. Monge, P.R., Contractor, N.S.: Theories of Communication Networks. Oxford University Press, Oxford (2003)
12. Carley, K.M.: An approach for relating social structure to cognitive structure. Journal of Mathematical Sociology **12(2)** (1986) 137–189
13. Carroll, J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In: Proceedings of the Fourteenth International World Wide Web Conference (WWW2005), Chiba, Japan. Volume 14. (2005) 613–622
14. Oegema, D., Kleinnijenhuis, J., Anderson, K.: Blaming and flaming: the influence of mass media content on interactions in on-line discussions. In Konijn, E.A., Tanis, M., Utz, S., eds.: Mediated Interpersonal Communication (forthcoming). Erlbaum, Mahwah, NJ (2006)
15. Kleinnijenhuis, J., Oegema, D., de Ridder, J., van Hoof, A., Vliegenthart, R.: De puinhopen in het nieuws. Volume 22 of Communicatie Dossier. Kluwer, Alphen aan de Rijn (Netherlands) (2003)
16. Dearing, J., Rogers, E.: Agenda setting. Sage, Thousand Oaks, CA (1996)
17. McCombs, M.E.: Setting the Agenda: The Mass Media and Public Opinion. Polity Press, Cambridge (2004)
18. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count. Lawerence Erlbaum Associates, Mahwah, NJ (2001)
19. Tateo, L.: The Italian extreme right on-line network: An exploratory study using an integrated social network analysis and content analysis approach. Journal of Computer-Mediated Communication **10**(2) (2005) article 10. http://jcmc.indiana.edu/vol10/issue2/tateo.html
20. Van Cuilenburg, J., Kleinnijenhuis, J., De Ridder, J.: Tekst en Betoog: naar een Computergestuurde Inhoudsanalyse van Betogende Teksten. Coutinho, Muiderberg (Netherlands) (1988)
21. Doreian, P., Batagelj, V., Ferligoj, A.: Generalized Blockmodeling. Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge (2005)
22. Van Atteveldt, W., Schlobach, S.: A modal view on polder politics. In: Proceedings of Methods for Modalities (M4M) 2005 (Berlin, 1-2 December). (2005)
23. Van Atteveldt, W., Schlobach, S.: Querying heterogeneous news repositories using RDF. Technical report, Free University Amsterdam (http://www.cs.vu.nl/ wva/pub/netrdf06.pdf) (2006)
24. Noy, N., Rector, A.: Defining n-ary relations on the semantic web. Working Draft for the W3C Semantic Web best practices group (2005)
25. MacGregor, R., Ko, I.Y.: Representing contextualized data using semantic web tools. In: Practical and Scalable Semantic Web Systems (workshop at second ISWC). (2003)

Measuring semantic centrality based on building consensual ontology on social network

Jason J. Jung and Jérôme Euzenat

INRIA Rhône-Alpes
ZIRST 655 avenue de l'Europe, Montbonnot,
38334 Saint Ismier cedex, France
j2jung@intelligent.pe.kr, Jerome.Euzenat@inrialpes.fr

Abstract. We have been focusing on three-layered socialized semantic space, consisting of social, ontology, and concept layers. In this paper, we propose a new measurement of semantic centrality of people, meaning the power of semantic bridging, on this architecture. Thereby, the consensual ontologies are discovered by semantic alignment-based mining process in the ontology and concept layer. It is represented as the maximal semantic substructures among personal ontologies of semantically interlinked community. Finally, we have shown an example of semantic centrality applied to resource annotation on social network, and discussed our assumptions used in formulation of this measurement.

1 Introduction

We have been focusing on constructing socialized semantic space to efficiently provide semantic collaboration and interoperability between people. With the emergence of semantic web, users (or actors) on social network have been applying their own personal ontologies to annotate the resources for improving interoperability between each other. However, as the number of users and ontologies are dramatically increasing, the structure of these networks are getting complex. Then, people are suffering from sharing and searching for the relevant information from the networks. In order to solve this problem, we have proposed a three-layered architecture for constructing socialized semantic space, (shown in Fig. 2) [1]. This space is designed to propagate the relational information not only within a layer but also between layers. We have provided the principles for extracting similarity between concepts and propagating this similarity to a distance and an alignment relation between ontologies.

In this paper, we define the notion of semantic centrality, which expresses the power of controlling *semantic* information flow on social network, and propose a novel network analysis method for measuring semantic centrality. Thereby, we need to discover the consensual ontology $\mathcal{CO}$ from personal ontologies applied to annotate the resources in personal information repositories. In fact, social network analysis (SNA) has regarded a consensus implying the central principles underlying the network as an important challenge [2]. With respect to semantic interoperability between heterogeneous information sources, consensual ontology is playing a role of a "semantic pivot" between heterogeneous information sources [3]. Here we assume that the consensual ontology should be simply organized as a set of concepts which are "most commonly" used in personal ontologies, as well as the relations among these concepts.

Basically, data mining methods are to uncover the hidden (more exactly, frequent) patterns from a given dataset like transactional databases. They also have shown the power of analyzing the structured datasets from various domains. Such datasets are not only XML documents [4] but also web link structure (or topology) [5], and protein structures [6].

In the similar way, we are motivated to extend simple frequent pattern mining method (e.g., *Apriori* algorithm) to *semantic substructure mining* (SSM) algorithm building consensus ontology, because ontologies are basically composed of a set of classes (or concepts) $\mathcal{C}$ and relations $\mathcal{R}$ between the classes [7]. In terms of social network, we can exploit the consensual ontology to measure semantic centrality of the participants on the corresponding social network, with respect to the quantity of major semantic information.

The remainder of the paper is organized as follows. Section 2 simply reviews several definitions in previous studies. Section 3 explains the semantic substructure mining algorithm for building consensual ontology, and then section 4 addresses the semantic centrality measurement from a given subgroup in social network. In section 5, we show an example and argue main contributions of this study by comparing with related work. Finally, section 6 will draw a conclusion and mention some issues as our future work.

2 Centrality measures on social network

A social network is denoted as a graph $G = (N, E)$, where N (a set of nodes) and E (a set of edges) represent users and links between users, respectively. In this paper, we consider only directed and labeled graphs. A path $p(i, j)$ between two arbitrary users u_i and u_j in graph G is a sequence of nodes and edges $\langle n_0, n_1\rangle, \langle n_1, n_2\rangle, \ldots, \langle n_{k-1}, n_k\rangle$, beginning with u_i ($= n_0$) and ending with u_j ($= n_k$), such that each edge connects its preceding with its succeeding node. The length of path is the number of edges (here, k), and we denote the set of shortest paths between u_i and u_j as $SP(i, j)$. Thus, the shortest path distance $spd(i, j)$ between two users u_i and u_j is the minimum element from $SP(i, j)$. Additionally, by Bellman criterion [8], let $\sigma_{i,j}(n)$ indicate the number of shortest paths $p(i, j) \in SP(i, j)$ that node $n \in N$ lies on. Basically, the centrality measures of a user are computed by using several features on the social network, and applied to determine the structural power. So far, in order to extract the structural information from a given social network, various measurements such as centrality [9], pair closeness [10], and authoritative [11] have been studied to realize the social relationships among a set of users. Especially, the centrality can be a way of representing the geometrical power of controlling *information flow* among participants on social network.

Table 1 shows four kinds of centrality measurements. Centrality C_C and C_G are based on the distances with the rest of nodes, while C_S and C_B emphasize the medium mediating between a pair of nodes. These are dependent upon the notion of the characteristics of social network. Also, we may apply hybrid approach of topological features, as combining different centrality measurements.

Table 1. Centrality measurements on social network

Closeness centrality C_C [12]	$C_C(n) = \frac{1}{\sum_{t \in N} spd(n,t)}$
Graph centrality C_G [9]	$C_G(n) = \frac{1}{\max_{t \in N} spd(n,t)}$
Stress centrality C_S [13]	$C_S(n) = \sum_{s \neq n \neq t \in N} \sigma_{s,t}(n)$
Betweenness centrality C_B [14]	$C_B(n) = \sum_{s \neq n \neq t \in N} \frac{\sigma_{s,t}(n)}{\lvert SPs,t \rvert}$

3 Discovering consensual ontology

In this section, we explain how to build the consensual ontology. Thereby, we focus on extracting the most frequent and common classes from a set of ontologies on social network. Substructure mining method will be briefly described, and then, we will show how it is applied to discover consensual ontology.

3.1 Background of substructure mining

Basically, data mining process (e.g., *Apriori* algorithm) can find out the correlation between items by statistical analysis of their occurrences in a given database. It consists of two steps; i) generating the candidate combinations, and ii) pruning by evaluating them with user-specified constraints like minimum support and confidence.

As extending to structured datasets, graph (or tree) mining is to discover the maximal frequent substructures from a given graph-structured dataset. For generating the candidates, the topological analysis is needed to justify whether each subgraph G' is a candidate of a given graph G or not. $G' = (N', E')$ is *induced* from $G = (N, E)$, represented as $G' \preceq G$, if and only if there exists a mapping function $\theta : N' \rightarrow N$ such that i) for each node $n' \in N'$, $n = \theta(n')$, and ii) for each edge $\langle n'_i, n'_j \rangle \in E'$, $\langle n_i, n_j \rangle = \langle \theta(n'_i), \theta(n'_j) \rangle$. Hence, only $G' \preceq G$ can be included in a set of candidate subgraphs [15].

Next, each candidate's support is given by $SUP(G') = \frac{Freq_D(G')}{|DB|}$ where $|DB|$ is the number of graphs in a given database DB, and $Freq_D(G') = \sum_{T \in DB} Occur(G')$ counting the frequency of subgraph G'. The G' of which support value is less than minimum support has to be discarded. The candidates over minimum support are joined each other to find out the larger subgraph. After repeating these steps, eventually, the maximal frequent subgraph can be uncovered.

Particularly, in subtree mining, *PatternMiner* and *TreeMiner* propose a level-wise algorithm based on *Apriori* scheme [16] for mining association rules and depth-first searching for using the novel scope-list, respectively [15].

3.2 Semantic substructure mining algorithm

The personal ontologies $\mathcal{PO}$ on social network is the target of this paper. We regard these personal ontologies as graph-structured knowledge, because they are generated by merging the ontology fragments derived from the reference ontologies by the corresponding user's manual (or semi-automatic) coding [1].

Given a set of personal ontologies, we focus on discovering the consensual ontology $\mathcal{CO}$ under *Apriori* assumption. As extending basic idea of data mining (described in the previous section), the semantic substructure mining algorithm SSM follows the three steps;

1. initialization of a set of candidate classes $CDT^1 = \{\ldots, \{c_i\}, \ldots\}$,
2. expansion of CDT^{t-1} to $\widetilde{CDT}^t = \{\ldots, \{c_i, \ldots, c_{i+t-1}\}, \ldots\}$ by join operation, and
3. refinement of CDT^t by evaluation with user-specific minimum support τ_{SUP}

where $\widetilde{CDT}$ and CDT indicate the power sets including the frequent class sets and the candidate class sets, respectively. The second and third steps are repeated until the constraints such as minimum supports are met ($t = T$). It means that we can finally get the consensual ontology which is composed of T classes.

Semantically induced substructure A candidate class is supposed to be a substructure *semantically induced* from the set of ontologies, and it is represented by

$$cdt_i^t \preceq^{\diamond} PO_k \iff SemInd(cdt_i^t, PO_k) \geq \zeta \tag{1}$$

where $cdt_i^t \in \widetilde{CDT}^t$ and $PO_k \in \mathcal{PO}$. For testing this induction, matching two ontologies has to be conducted by using the semantic similarity measurement, proposed in [17], rather than simple string-matching, in order to reduce some lexical heterogeneity problems such as synonyms. Hence, $SemInd$ is given by

$$SemInd(o_i, O) = \max \frac{\sum_{\langle c,c' \rangle \in Pairing(o_i,O)} Sim_C(c, c')}{|o_i|} \tag{2}$$

in which $Pairing$ provides a matching of the two set of classes. It is established by finding the best matching which is maximizing the summation of the similarities between the classes. The basic notion can be described that two entities are more similar, if they have the more similar features. Then, the class similarity measure Sim_C is formulated as

$$\begin{aligned} Sim_C(c, c') &= \sum_{E \in \mathcal{N}(C)} \pi_E^C MSim_Y(E(c), E(c')) && (3) \\ &= \pi_L^C Sim_L(c, c') + \pi_{sup}^C MSim_C(E^{sup}(c), E^{sup}(c')) \\ &+ \pi_{sub}^C MSim_C(E^{sub}(c), E^{sub}(c')) \\ &+ \pi_{sib}^C MSim_C(E^{sib}(c), E^{sib}(c')) && (4) \end{aligned}$$

where $\mathcal{N}(C) \in \{E^1, \ldots, E^n\}$ is the set of all relationships in which classes are involved (in this paper, we are considering three relationships; superclass, subclass, and sibling class), and π^C is the normalized weighting factor to the corresponding relationships. Also, similar to Eq. 2, the set function $MSim_C$ is given by

$$MSim_C(S, S') = \frac{\max \sum_{\langle c,c' \rangle \in Pairing(S,S')} Sim_C(c, c')}{\max(|S|, |S'|)}. \tag{5}$$

Finally, $SemInd(cdt_i^t, PO_k)$ is assigned into $[0, 1]$. Thus, cdt_i^t of which similarity with a given ontology PO_k is over ζ is regarded as one of semantically induced substructures from PO_k. When $\zeta = 1$, only candidates exactly matched will be chosen without concerning about the semantic heterogeneity.

Expansion and refinement by evaluation In order to discover the maximal frequent substructure, we have to repeat these two processes; expansion for generating candidates and refinement. Refinement process of candidates induced from personal ontologies, exactly same as in general data mining, is to compare the frequency of the corresponding substructure candidate with user-specific threshold (e.g., minimum support τ_{SUP}). The candidate cdt_i^t extracted through comparing the similarities measured by $SemInd$ with ζ can be counted as the occurrence in the set of personal ontologies $\mathcal{PO}$. Function $Occur^{\diamond}$ returns 1, if $cdt_i^t \preceq^{\diamond} PO_k$. Otherwise, it returns 0. Thus, frequency of a candidate is $Freq_{\mathcal{PO}}(cdt_i^t) = \sum_{PO_k \in \mathcal{PO}} Occur^{\diamond}(cdt_i^t)$, and the support is given by

$$SUP(cdt_i^t) = \frac{Freq_{\mathcal{PO}}(cdt_i^t)}{|\mathcal{PO}|} = \frac{\sum_{PO_k \in \mathcal{PO}} Occur^{\diamond}(cdt_i^t)}{|\mathcal{PO}|}. \tag{6}$$

Only the candidate set of classes cdt_i^t of which support $SUP(cdt_i^t) \geq \tau_{SUP}$ can be chosen to generate the expanded candidates $\widetilde{CDT}^{t+1}$.

After a set of candidate features CDT^1 is initially selected by

$$CDT^1 = \{cdt_i^1 | SUP(cdt_i^1) \geq \tau_{SUP}\}, \tag{7}$$

we have to expand the set of candidate class sets and refine them where $t \geq 2$. Thus, CDT^t is obtained by

$$CDT^t = refine(\widetilde{CDT}^t) \tag{8}$$

$$= refine(expand(CDT^{t-1})) \tag{9}$$

where function $refine$ is to evaluate $\begin{pmatrix} |CDT^{t-1}| \\ t \end{pmatrix}$ set elements generated by function $expand$ where $|CDT^{t-1}|$ is the total number of the single classes in CDT^{t-1}.

3.3 Consensual ontology and semantic subgroup discovery

By using semantic substructure mining algorithm, the maximal semantic substructures were able to be obtained from a given set of personal ontologies. Then, the consensual ontology $\mathcal{CO}$ is represented as $\{cdt_i^T | cdt_i^T \in CDT^T, SUP(cdt_i^T) \geq \tau_{SUP}\}$ when $\widetilde{CDT}^{T+1}$ is an empty set.

However, we have to realize the problem when the target social network is intermingled with semantically heterogeneous communities. Substructure mining algorithm based on counting simple occurrence (or frequency) analysis is difficult to build more than two consensual ontologies at the same time. Thereby, the social network should be fragmented into the communities (or groups [18]) whose semantic preferences are more

cohesive with each other than others. In other words, this is similar to *user clustering* based on the semantic cohesion among users on the social network. Thus, let K the number of communities (user groups) on social network. The best combination of user groups is obtained by maximizing the objective function $F_{SubGroup}(UG_1, \ldots, UG_K)$

$$\max F_{SubGroup} = \max \frac{\sum_{k=1}^{K} Distance(\mathcal{CO}_i, \mathcal{CO}_j)}{K} \tag{10}$$

$$= \max \frac{\sum_{k=1}^{K} (1 - Sim_C(c \in \mathcal{CO}_i, c' \in \mathcal{CO}_j))}{K} \tag{11}$$

$$\approx \min \frac{\sum_{k=1}^{K} Sim_C(c \in \mathcal{CO}_i, c' \in \mathcal{CO}_j)}{K} \tag{12}$$

where $\mathcal{CO}_i = SSM(UG_i)$. Function $Distance$ is derived from similarity measure by taking its complement to 1. Through this equation, the underlying communities can be found out. Each time the function $refine$ of SSM algorithm is finished, this process should be conducted.

4 Semantic centrality

As mentioned in Table 1, there have been several centrality indices to measure the power of structural position on social network. However, they are not appropriate to reflect the centrality among the underlying semantic relationships between personal ontologies on the socialized semantic network introduced in our previous work [1].

We define a semantic centrality as the power of semantic bridging on the semantic social network. Suppose that two users s and t are not able to communicate with each other, due to the semantic heterogeneity between their personal ontologies PO_s and PO_t. Thereby, we need to search for the personal ontology PO_i of which semantic centrality is high enough to reconcile these ontologies. It means PO_i is containing some classes matched with the consensual ontology $\mathcal{CO}$. We intuitively assume that a user is assigned higher semantic centrality, as his personal ontology includes more consensual classes in common. Thus, we formulate a semantic centrality of i-th user $C^{\diamond}(i)$ as

$$C^{\diamond}(i) = \frac{|PO_i \cap \mathcal{CO}|}{|PO_i|} \sum_{s \neq t \neq i \in N} \frac{\sigma^{\diamond}_{PO_s, PO_t}(PO_i)}{|SP^{\diamond}(s,t)|} \tag{13}$$

which means the semantic closeness (or coverage) of the personal ontology PO_i to the discovered consensual ontology $\mathcal{CO}$. The denominator $|PO_i|$ is for the normalization by the total number of classes organizing the personal ontology. $SP^{\diamond}$ is a pair of users whose personal ontologies are not semantically interoperable directly. So, C_B can be replaced by C_C or others. More importantly, function $\sigma^{\diamond}$ is to determine the efficiency of reconciliation, and it is given by

$$\sigma^{\diamond}_{PO_s, PO_t}(PO_i) = \frac{|PO_s \cap PO_i| \cdot |PO_t \cap PO_i|}{|PO_s \cap \mathcal{CO}| \cdot |PO_t \cap \mathcal{CO}|} \tag{14}$$

which expresses that the number of matched classes between two ontologies is in linear proportion, in contrast of that of matched classes with consensus ontologies. Additionally, in Eq. 13 and 14, the counting computation of union sets is done by

$$|A \cap B| = count(\langle c, c' \rangle)_{\langle c,c' \rangle \in Pairing(A,B), Sim_C(c,c')=1}. \tag{15}$$

As next issue, we note that there are two kinds of semantic centrality measurements, with respect to the scope and the topologies of communities.

- *Local* semantic centrality $C_L^\diamond$ means the power of semantic bridging between the members within the same community.
- *Global* semantic centrality $C_G^\diamond$ implies the measurement of the bridging power between two communities.

Then, for computing these centrality measurements, the communities on the whole social network should be firstly organized by using SSM algorithm. Local semantic centrality $C_L^\diamond$ is computed by

$$C_L^\diamond(i) = \alpha_T C_B(i) + (1 - \alpha_T) C^\diamond(i) \tag{16}$$

where the first term is for reflecting the effect of physical (or explicit) social linkage of a given community (mentioned in Table 1), and the second term is semantic centrality $C^\diamond$. The coefficient $\alpha_T \in [0, 1)$ is to control the portion of topological effects. This is formulated as linearly combined with topological centrality measurements and semantic centrality in Eq. 13.

On the other hand, global semantic centrality $C_G^\diamond(i, X)$ of i-th user to a certain community X is based on three factors; i) topologically, the betweenness centrality between the people of two communities I, including i-th user, and X, ii) the similarity between the consensual ontology of target communities X and the corresponding personal ontology PO_i, and iii) the corresponding local semantic centrality. Thus, as linearly combined, it is given by

$$\begin{aligned} C_G^\diamond(i, X) = \beta_T C_B(i) & \frac{\sum_{x \in X} link(i, x)}{\sum_{u \in I, x \in X} link(u, x)} \\ & + \beta_S MSim_C(PO_i, CO_X) \\ & + \beta_{LS} C_L^\diamond(i) \sum_{j \in X, link(i,j)=1} C_L^\diamond(j) \end{aligned} \tag{17}$$

where $\beta_T, \beta_S, \beta_{LS} \in [0, 1)$ are the coefficients controlling the portion of topological effects, similar to α_T, the similarity effects, and local semantic centrality effects, respectively. For normalization, $\beta_T + \beta_S + \beta_{LS} = 1$. First term simply indicates the ratio of the linkages by the corresponding user to the total linkages with the target community. User j in the target community X is a member linking to the i-th user. In second term, we put the ontology similarity between consensual ontology of community X and personal ontology, because the more similar classes make the mediation powerful. Finally, the third term applies the local semantic centrality. When a user are linked with more "semantically" central users in a community, his global centrality becomes increased.

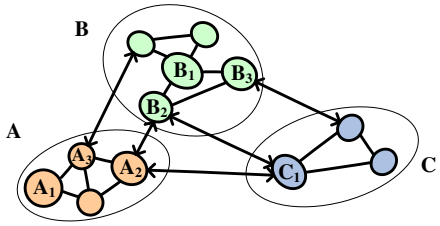


Fig. 1. Two kinds of semantic centrality measurements

We want to show two cases, with respect to $C_L^\diamond$ and $C_G^\diamond$. Let three communities (A, B, and C) organized from a given social network (the number of communities $K = 3$ in Eq. 12), as shown in Fig. 1. We assume that user A_1, B_1, and C_1 are is the highest $C_L^\diamond$ within each community. First case, most possibly, is that the user whose local semantic centrality is highest is also assigned the highest global semantic centrality, e.g., $C_G^\diamond(C_1, B)$ (or $C_G^\diamond(C_1, A)$) calculated as

$$C_G^\diamond(C_1, B) = \beta_T(\frac{1}{3})(\frac{2}{2}) + \beta_S MSIM_C(CO_B, PO_{C_1}) + \beta_{LS} C_L^\diamond(C_1)\{C_L^\diamond(B_2) + C_L^\diamond(B_3)\}$$

where $C_L^\diamond(C_1)$ is assumed to be larger than any other members in community C. Moreover, topologically, C_1 is the only channel to communicate with other communities. As second case, in community B to C, even though B_1's local centrality is the highest, $C_G^\diamond(B_2, C)$ might be higher due to the linkage patterns with C. This is also larger than $C_G^\diamond(B_3, C)$, because C_1 is assigned the highest local semantic centrality.

Here, the scenario is given for explaining how semantic centrality is applied to. Above all, Fig. 2 shows the three-layered architecture of semantic social network, which is composed of a social layer, an ontology layer, and a concept layer. While social layer can simply store the physical connections between users, ontology layer represents the personal ontologies applied to annotate the corresponding user's resources. The classes organizing these ontologies are located in concept layer, so that they are aligned to measure the similarity. Assume that five users are organizing social network and their links are shown in social layer. With respect to the traditional centrality measurements, we

Table 2. Measuring centrality on social network

Centrality	Antoine	Jerome	Jason	Arun	Sebastien
Closeness centrality	1/7	1/6	**1/5**	1/7	1/9
Betweenness centrality	4/10	7/10	**9/10**	4/10	4/10
Semantic centrality		$\star$			

found out that *Jason* is the most powerful user, as shown in Table 2. However, during

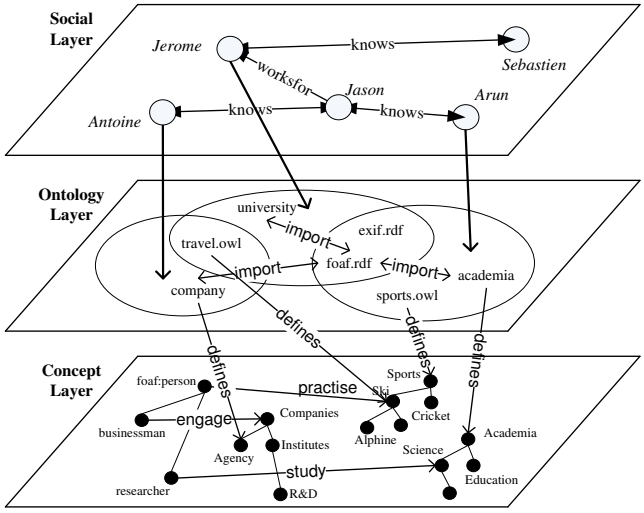


Fig. 2. A three-layered social semantic network

bottom-up inference (which is from concept layer to social layer), the underlying hidden patterns are uncovered. The semantic bridging between *Antoine* and *Arun* is provided only *Jerome*'s personal ontology. People have been trying to annotate the resources (e.g., photos) in their own repository by using personal ontologies. In order to increase the efficiency of annotation, they have been needed to share the semantic information between each other or photos themselves. The problem is caused by semantic heterogeneity of annotations, e.g., between *Antoine* and *Arun*. Thereby, they should search for the users who can most likely play a role of a semantic bridge between them, and ask him (e.g., *Jerome*) to translate the annotations for mutual understanding.

Next, we can consider semantic centrality to propagate semantic information. The semantic information can be newly asserted or changed over time, and it is supposed to be announced to all other users on social network. For the purpose of efficient network management, up-to-date semantic information can be propagated in order of semantic centrality. Somehow, *Arun* acquires some new information and update his personal ontology. Thus, this event has to be notified to *Jerome*, rather to *Jason*, even though he is directly linked with *Jason*.

5 Discussion and related work

Many systems have been interested in information sharing on the distributed systems. With emergent of semantic web environment, rather than content, semantic information has been the target data to be exchanged. Since EDUTELLA introduced an infrastructure for exchanging metadata on peer-to-peer (P2P) environment [19], several querying-based systems have been implemented such as Bibster [20], Oyster [21], and SQAPS [22]. Our goal is also to search for relevant semantic information and share it with other

users on distributed information space, but we are more focusing on how peers are interlinked, the so-called social network analysis. Especially, as the contributions of this paper, we proposed two main ideas;

1. consensual ontology discovery by semantic substructure mining algorithm and
2. local versus global semantic centrality measurement.

Firstly, in order to construct consensual ontologies, this paper introduced a pruning-based approach to discover the maximal frequent substructure. In contrast, as the most general approach, Stephens et al. have organized the consensual (or global) ontologies by exhaustive merging of a given set of ontologies [3] in order to retain or maximize the chance to be semantically bridged. Additionally, [23] mentions ontology-based consensus making process from the user communities. It also seems to try to find out the consensus ontology, after organizing the communities from people.

Meanwhile, in order to support the communication between communities, in [24], Breslin et al. proposed the SIOC (Semantically Interlinked Online Communites) ontology[1], rather than discovering the consensus.

Especially, in terms of the efficiency of query propagation on peer-to-peer environment, several studies have introduced the systems based on semantic overlay network (SON). They are based on the broadcasting scheme. After the queries are semantically analyzed, the relevant topics are distilled. Through the multiple overlay network, the queries are propagated to either the set of selected nodes [25], or the super-peers [26]. In contrast, in our method, the queries should be sent to the node whose semantic centrality is largest.

Finally, we want to discuss the personal ontology built by people. In this study, the personal ontology is assumed to play a role of the important evidence reflecting the preferences of corresponding user. However, because users can refer to the upper-level ontologies and even import the other user's personal ontologies, it is too ambiguous to measure the similarity between personal ontologies for the consensus.

6 Concluding remarks and future work

As a conclusion, we put forward a new measurement for semantic centrality, expressing the potential power of semantic bridging among users on social network. Consensual ontologies thereby were built by semantic substructure mining algorithm, and they had the capability to discover the subgroups whose semantic preferences are relatively closer than others. More importantly, the notion of them was designed to be adaptable to the three-layered architecture (social, ontology, and concept layer) for socialized semantic space [1]. The three-layered architecture provides two ways of inference for the hidden relationships between entities; top-down (from social layer to concept layer) and bottom-up (reversely). This paper is related to the bottom-up inference. Especially, we want to mention that the communities on social network are organized with respect to *semantic preference* implicitly reflected during designing and using their own personal ontologies.

[1] SIOC. http://rdfs.org/sioc/

For dealing with this problem mentioned in Sect. 5, we have to track the user actions and interactions. For instance, similar to [27], we may consider only the concepts applied to the specific resources. On the other hand, the "concepts with dust," which is not used for a long time, should be degraded by using machine learning methodologies.

As future work of semantic centrality, we have three main plans to investigate the followings issues

- *semantic subgroup discovery*, to organize the sophisticated user groups with enhancing Eq. 12,
- *query propagation*, to determine the ordering (or route) of potential peers to which the queries will be sent, and
- *semantic synchronization*, to maximize the efficiency interoperability by information diffusion.

Furthermore, we have to consider to enhance the semantic centrality measurement $C^{\diamond}$ by combining with i) authoritative and hub centrality measurement, proposed in [11], and ii) the modified shortest paths $spd(n,t) = \frac{1}{C^{\diamond}(n)+C^{\diamond}(t)}$. Finally, like [28], we have plan to visualize the semantic dynamics on the social network.

References

1. Jung, J.J., Euzenat, J.: From personal ontologies to semantic social space. In: Poster of the 4th European Semantic Web Conference (ESWC 2006). (2006)
2. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1994)
3. Stephens, L.M., Gangam, A.K., Huhns, M.N.: Constructing consensus ontologies for the semantic web: A conceptual approach. World Wide Web **7**(4) (2004) 421–442
4. Braga, D., Campi, A., Klemettinen, M., Lanzi, P.L.: Mining association rules from XML data. In Kambayashi, Y., Winiwarter, W., Arikawa, M., eds.: Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2002), Aix-en-Provence, France. Volume 2454 of Lecture Notes in Computer Science., Springer (2002) 21–30
5. Chakrabarti, S., Dom, B., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., Kleinberg, J.M.: Mining the web's link structure. IEEE Computer **32**(8) (1999) 60–67
6. Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J., Tropsha, A.: Mining protein family specific residue packing patterns from protein structure graphs. In: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology (RECOMB '04), New York, NY, USA, ACM Press (2004) 308–315
7. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition **5**(2) (1993) 199–220
8. Brandes, U.: A faster algorithm for betweenness centrality. Journal of Mathematical Sociology **25**(2) (2001) 163–177
9. Haga, P., Harary, F.: Eccentricity and centrality in networks. Social Networks **17**(1) (1995) 57–63
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences **99** (2002) 7821–7826
11. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of ACM **46**(5) (1999) 604–632
12. Sabidussi, G.: The centrality index of a graph. Psychometirka **31** (1966) 581–603

13. Shimbel, A.: Structural parameters of communication networks. Bulletin of Mathematical Biophysics **15** (1953) 501–507
14. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40** (1977) 35–41
15. Zaki, M.J.: Efficiently mining frequent trees in a forest: Algorithms and applications. IEEE Transactions on Knowledge and Data Engineering **17**(8) (2005) 1021–1035
16. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94), San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1994) 487–499
17. Euzenat, J., Valtchev, P.: Similarity-based ontology alignment in OWL-Lite. In de Mántaras, R.L., Saitta, L., eds.: Proceedings of the 16th European Conference on Artificial Intelligence (ECAI'2004), Valencia, Spain, August 22-27, 2004, IOS Press (2004) 333–337
18. Kleinberg, J.M.: Small-world phenomena and the dynamics of information. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada], MIT Press (2001) 431–438
19. Nejdl, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmér, M., Risch, T.: EDUTELLA: a P2P networking infrastructure based on rdf. In: Proceedings of the 11th international conference on World Wide Web (WWW '02), New York, NY, USA, ACM Press (2002) 604–615
20. Haase, P., Broekstra, J., Ehrig, M., Menken, M., Mika, P., Olko, M., Plechawski, M., Pyszlak, P., Schnizler, B., Siebes, R., Staab, S., Tempich, C.: Bibster - a semantics-based bibliographic peer-to-peer system. In McIlraith, S.A., Plexousakis, D., van Harmelen, F., eds.: Proceedings of the Third International Semantic Web Conference (ISWC 2004),Hiroshima, Japan, November 7-11, 2004. Volume 3298 of Lecture Notes in Computer Science. (2004) 122–136
21. Palma, R., Haase, P.: Oyster - sharing and re-using ontologies in a peer-to-peer community. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: Proceedings of the 4th International Semantic Web Conference (ISWC 2005), Galway, Ireland, November 6-10, 2005. Volume 3729 of Lecture Notes in Computer Science., Springer (2005) 1059–1062
22. Fernandez-Garcia, N., Sanchez-Fernandez, L., Blazquez, J.M., Larrabeiti, D.: An ontology-based P2P system for query-based semantic annotation sharing. In: Proceedings of Workshop on Ontologies in Peer-to-Peer Communities at the ESWC 2005, May 30, Heraklion, Greece. (2005)
23. Zhdanova, A.V., Martín-Recuerda, F.: Consensus making on the semantic web: Personalization and community support. In Ngu, A.H.H., Kitsuregawa, M., Neuhold, E.J., Chung, J.Y., Sheng, Q.Z., eds.: Proceedings of the 6th International Conference on Web Information Systems Engineering (WISE 2005), November 20-22, 2005. Volume 3806 of Lecture Notes in Computer Science., Springer (2005) 599–600
24. Breslin, J.G., Harth, A., Bojars, U., Decker, S.: Towards semantically-interlinked online communities. In Gómez-Pérez, A., Euzenat, J., eds.: Proceedings of the Second European Semantic Web Conference (ESWC 2005), May 29 - June 1, 2005. Volume 3532 of Lecture Notes in Computer Science., Springer (2005) 500–514
25. Crespo, A., Garcia-Molina, H.: Semantic overlay networks for p2p systems. In Moro, G., Bergamaschi, S., Aberer, K., eds.: Proceedings of the 3rd International Workshop Agents and Peer-to-Peer Computing (AP2PC 2004), July 19, 2004. Volume 3601 of Lecture Notes in Computer Science., Springer (2005) 1–13
26. Löser, A., Naumann, F., Siberski, W., Nejdl, W., Thaden, U.: Semantic overlay clusters within super-peer networks. In Aberer, K., Kalogeraki, V., Koubarakis, M., eds.: Proceedings of the First International Workshop on Databases, Information Systems, and Peer-to-Peer Computing (DBISP2P), September 7-8, 2003. Volume 2944 of Lecture Notes in Computer Science., Springer (2004) 33–47

27. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: Proceedings of the 4th International Semantic Web Conference (ISWC 2005), November 6-10, 2005. Volume 3729 of Lecture Notes in Computer Science., Springer (2005) 522–536
28. Jung, J.J.: Visualizing recommendation flow on social network. Journal of Universal Computer Science **11**(11) (2005) 1780–1791

Building Emergent Social Networks and Group Profiles by Semantic User Preference Clustering

Iván Cantador, Pablo Castells

Escuela Politécnica Superior, Universidad Autónoma de Madrid
Campus de Cantoblanco, 28049 Madrid, Spain
{ivan.cantador, pablo.castells}@uam.es

Abstract. This paper presents a novel approach to automatic semantic social network construction based on semantic user preference clustering. Considering a number of users, each of them with an associated ontology-based profile, we propose a strategy that clusters the concepts of the reference ontology according to user preferences of these concepts, and then determines which clusters are more appropriate to the users. The resultant user clusters can be merged into individual group profiles, automatically defining a semantic social network suitable for use in collaborative and recommendation environments.

1 Introduction

The swift development, spread, and convergence of information and communication technologies and support infrastructures, reaching all aspects of businesses and homes in our everyday lives, is giving rise to new and unforeseen ways of inter-personal connection, communication, and collaboration. Virtual communities, computer-supported social networks, and collective interaction are indeed starting to proliferate and grow in increasingly sophisticated ways, opening new opportunities for research on social group analysis, modeling, and exploitation. In this paper we propose a novel approach towards building emerging social networks by analyzing the individual motivations and preferences of users, broken into potentially different areas of personal interest.

Finding hidden links between users based on the similarity of their preferences or historic behavior is not a new idea. In fact, this is the essence of the well-known collaborative recommender systems (e.g. see [2,10,15]). However, in typical approaches, the comparison between users is done globally, in such a way that partial, but strong and useful similarities may be missed. For instance, two people may have a highly coincident taste in cinema, but a very divergent one in sports, or totally different professional interests. The opinions of these people on movies could be highly valuable for each other, but risk to be ignored by many collaborative recommender systems, because the global similarity between the users is low.

In this paper we propose a multi-layered approach to dynamic social networking. Like in previous approaches [1,13,14], our method builds and compares profiles of user interests for semantic topics and specific concepts, in order to find similarities among users. But in contrast to prior work, in our approach user profiles are divided

into clusters of cohesive interests, and based on this, several layers of networks are found. This provides a richer, finer-grained model of interpersonal links, which better represents the way people find common interests in real life, which typically takes place on different, partial planes of each other's life.

Our approach is based on an ontological representation of the domain of discourse where user interests are defined. The ontological space takes the shape of a semantic network of interrelated domain concepts. Taking advantage of the relations between concepts, and the (weighted) preferences of users for the concepts, our system clusters the semantic space based on the correlation of concepts appearing in the preferences of individual users. After this, user profiles are partitioned by projecting the concept clusters into the set of preferences of each user. Then, users can be compared on the basis of the resulting subsets of interests, in such a way that several, rather than just one, (weighted) links can be found between two users.

Multi-layered social networks are potentially useful for many purposes. For instance, users may share preferences, items, knowledge, and benefit from each other's experience in focused or specialized conceptual areas, even if they have very different profiles as a whole. Such semantic subareas need not be defined manually, as they emerge automatically with our proposed method. Users may be recommended items or direct contacts with other users for different aspects of day-to-day life.

In addition to these possibilities, we have experimented with the proposed two-way space clustering mechanism. Finding clusters of users based on those clusters of concepts that represent common topics of interest, we obtain a reinforced partition of the user space that can be exploited to build group profiles for sets of related users. These group profiles enable an efficient strategy for collaborative recommendation in real-time, by using the merged profiles as representatives of classes of users.

The rest of the paper has the following organization. Section 2 describes our ontology-based user profile representation and gives an overview of the personalized content retrieval system in which it is being used. Section 3 explains our proposal to automatic construction of multi-layered social networks based on semantic user preference clustering. In section 4 several strategies for modeling group profiles are experimentally investigated. Finally, some conclusions and future research lines are given in section 5.

2 User Profile Representation

Our research builds upon an ontology-based personalization framework. In this section we provide an overview of the basic principles of this framework, with a special focus on user profile representation, and the exploitation of the profiles for personalized content retrieval. Further details can be found in [16].

In contrast with other approaches in personalized content retrieval, our approach makes use of explicit user profiles (as opposed to e.g. sets of preferred documents). Working within an ontology-based personalization framework, in which the domain of interest is described through semantic concepts corresponding to the different classes and instances of a domain ontology, user preferences are represented as vectors

$U_m = (w_{m1},...,w_{mN})$, where $m = 1,...,M$, M is the number of existent user profiles, and $w_{mn} \in [0,1]$ is the weight that measure the intensity of the interest of user *m* for concept c_n in the domain ontology, *N* being the total number of concepts in the ontology. Similarly, the objects d_k in the retrieval space are assumed to be described (annotated) by vectors $\vec{c}_n = (c_{n1},...,c_{nM})$ of concept weights, in the same vector-space as user preferences. Comparing the metadata of content items, and the preferred concepts in a user profile, the system finds how the user may like each element. Based on her preference weights, measures of user interest for content units can be computed, with which it is possible to prioritize, filter and rank contents (a collection, a catalog section, a search result) in a personal way.

Ontology-based representations [12] are richer, more precise, less ambiguous than keyword-based or item-based models. They provide an adequate grounding for the representation of coarse to fine-grained user interests (e.g. interest for individual items such as a sports team, an actor, a stock value) in a hierarchical way, and can be a key enabler to deal with the subtleties of user preferences. An ontology provides further formal, computer-processable meaning on the concepts (who is coaching a team, an actor's filmography, financial data on a stock), and makes it available for the personalization system to take advantage of. Furthermore, ontology standards, such as RDF and OWL, support inference mechanisms that can be used in the system to further enhance personalization, so that, for instance, a user interested in animals (superclass of *cat*) is also recommended items about cats. Inversely, a user interested in *lizards*, *snakes*, and *chameleons* can be inferred to be interested in *reptiles* with a certain confidence. Also, a user keen of *Madrid* can be assumed to like *Spain*, through the *locatedIn* relation.

3 Emergent Semantic Social Networks

As explained above, our ontology-based personalization framework makes use of explicit user profiles. The users preferences are represented as vectors $U_m = (w_{m1},...,w_{mN})$, where the weights $w_{mn} \in [0,1]$ measure the intensity of the *m*-th user interest for each of the N concepts in the domain ontology. The weights thus represent a way of connecting the concept and the user preferences spaces (top left picture of Figure 1).

We propose here to exploit the links between users and concepts to extract relations among users and derive semantic social networks according to common interests. Analyzing the structure of the domain ontology and taking into account the semantic preference weights of the user profiles we shall cluster the domain concept space generating groups of interests shared by certain users. Thus, those users who share interests of a specific concept cluster will be connected in the network, and their preference weights will measure the degree of membership to each cluster.

The next subsections explain in more detail the steps followed in the clustering process, which is shown in Figure 1. An example will be given afterwards to illustrate our proposal.

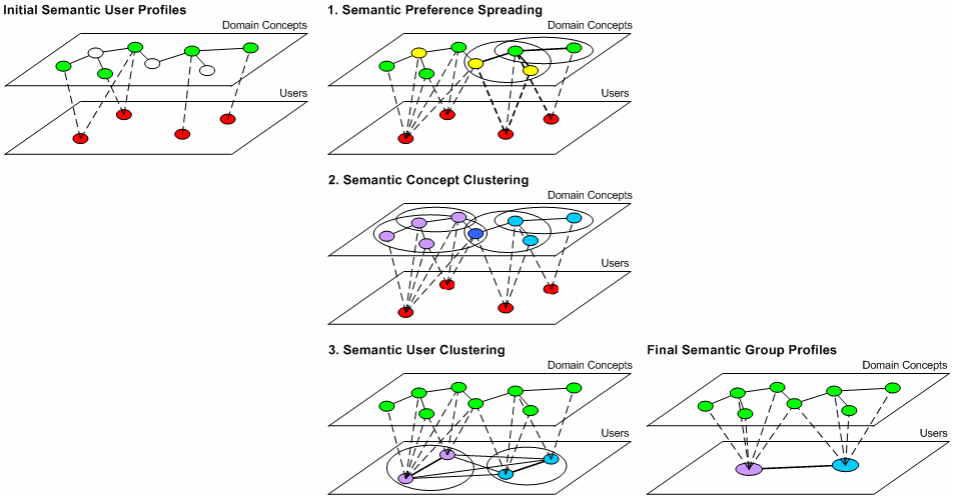


Fig. 1. Overall sequence of our proposed approach, comprising three steps: 1) semantic user preferences are spread, extending the initial sets of individual interests, 2) semantic domain concepts are clustered into concept groups, based on the vector space of user preferences, and 3) users are clustered in order to identify the closest class to each user

3.1 Semantic Preference Extension

In real scenarios, user profiles tend to be very scattered, especially in those applications where user profiles have to be manually defined. Users are usually not willing to spend time describing their detailed preferences to the system, even less to assign weights to them, especially if they do not have a clear understanding of the effects and results of this input. On the other hand, applications where an automatic preference learning algorithm is applied tend to recognize the main characteristics of user preferences, thus yielding profiles that may entail a lack of expressivity. To overcome this problem, we propose a semantic preference spreading mechanism, which expands the initial set of preferences stored in user profiles through explicit semantic relations with other concepts in the ontology (see picture numbered 1 in Figure 1). Our approach is based on the Constrained Spreading Activation (CSA) strategy [1,4,5]. The expansion is self-controlled by applying a decay factor to the intensity of preference each time a relation is traversed. For example, if an initial profile has a preference about *animals* with a weight of 0.7, the semantic CSA might add to the profile concepts such as *mammals* or *dog*, both of them with associated weights less than 0.7.

We have conducted experiments showing that the performance of the personalization system is considerably poorer when the spreading mechanism is not enabled. Typically, the basic user profiles without expansion are too simple. They provide a good representative sample of user preferences, but do not reflect the real extent of user interests, which results in low overlaps between the preferences of different users. Therefore, the extension is not only important for the performance of individual per-

sonalization, but is essential for the clustering strategy described in the following sections.

The enhancements achieved by the automatic preference extension mechanism show the benefits of an ontology-based representation of user profiles, in contrast to traditional, less expressive ones based on keywords and/or thematic categories.

3.2 Semantic Concept Clustering

In social communities it is fairly accepted that people who are known to share a specific interest are likely to have additional connected interests [9]. For instance, it is easy to understand that, in general, people who like *climbing*, also like topics related to *mountains* or topics related to other *adventure sports*. In fact, this assumption is the basis of most of the existing recommender and collaborative filtering recommender systems [2,10,15]. Here we take into account this hypothesis in order to cluster the concept space in groups of preferences shared by a number of users.

Specifically, for each concept c_n present in at least one of the *M* considered user profiles a vector $\vec{c}_n = (c_{n1},...,c_{nM})$ is assigned, where the component c_{nm} is the weight of concept c_n in the semantic profile of the *m*-th user or 0 if the concept does not appear on it. With these vectors a classical hierarchical clustering strategy [6] is applied. The obtained clusters (picture numbered 2 in Figure 1) thus represent in the concept-user vector space those groups of preferences (topics of interests) shared by the users.

Of course, several issues need to be addressed for the clustering algorithm, such as the distance measure between concepts and clusters, or the appropriate number of final clusters. These will be refined in future work. Here, as we shall explain in section 3.4, we have experimented with a simple example in which the number of clusters is known, and where we have used the Euclidean distance to measure the distances between concepts and an average linkage to measure the distances between clusters [6].

3.3 Semantic User Clustering

Co-clustering based recommender systems involve simultaneous clustering of users and items and generating predictions based on the average ratings of the generated co-clusters [3,8]. Following this idea, once the semantic concept clusters are created, we assign each user to a specific cluster. The similarities between a certain user profile $U_m = (w_{m1},...,w_{mN})$ and the different clusters C_k are computed by the following expression:

$$similarity(U_m, C_k) = \frac{\sum_{n:c_{nm} \in C_k} w_{nm}}{|C_k|} \quad \textbf{(1)}$$

where c_{mn} represents the concept associated to the *n*-th component of the user profile U_m, and $|C_k|$ is the number of concepts included in cluster C_k. The clusters with

highest similarities are then assigned to the users, thus creating groups of users with shared interests (picture numbered 3 in Figure 1).

The obtained user clusters can be used to define underlying semantic social networks. The preference weights of user profiles, the degrees of membership of the users to each cluster and the similarity measures between clusters provide mechanisms to describe the relations between two distinct types of social items: individuals and groups of individuals. As an applicative development of the obtained user semantic relations, in section 4 we give a first contribution investigating strategies for merging user profiles with common preferences to generate semantic group profiles (picture on the bottom right side of Figure 1). But before that, in the next subsection we describe an artificial experiment that shows an example of evolution and results generated from the presented clustering proposal.

3.4 A simple experiment

In order to check the feasibility of the explained clustering strategy an artificial problem has been set up for this work. The scenario of the problem is the following. A set of twenty user profiles are considered. Each profile is manually defined taking into account six possible topics: *motor*, *construction*, *family*, *animals*, *beach* and *vegetation*. The degree of interest each user has for the different topics are shown in Table 1.

Table 1. Degrees of interest of each user about the six considered topics, and expected user clusters to be obtained with our semantic preference clustering strategy

	Motor	***Construction***	***Family***	***Animals***	***Beach***	***Vegetation***	**Expected Cluster**
User1	High	High	Low	Low	Low	Low	*1*
User2	High	High	Low	Medium	Low	Low	*1*
User3	High	Medium	Low	Low	Medium	Low	*1*
User4	High	Medium	Low	Medium	Low	Low	*1*
User5	Medium	High	Medium	Low	Low	Low	*1*
User6	Medium	Medium	Low	Low	Low	Low	*1*
User7	Low	Low	High	High	Low	Medium	*2*
User8	Low	Medium	High	High	Low	Low	*2*
User9	Low	Low	High	Medium	Medium	Low	*2*
User10	Low	Low	High	Medium	Low	Medium	*2*
User11	Low	Low	Medium	High	Low	Low	*2*
User12	Low	Low	Medium	Medium	Low	Low	*2*
User13	Low	Low	Low	Low	High	High	*3*
User14	Medium	Low	Low	Low	High	High	*3*
User15	Low	Low	Medium	Low	High	Medium	*3*
User16	Low	Medium	Low	Low	High	Medium	*3*
User17	Low	Low	Low	Medium	Medium	High	*3*
User18	Low	Low	Low	Low	Medium	Medium	*3*
User19	Low	High	Low	Low	Medium	Low	*1*
User20	Low	Medium	High	Low	Low	Low	*2*

For a certain user and a certain topic, a *high* degree of interest means that the user semantic profile has weights close to 1 in some of the concepts corresponding to the topic, a *medium* degree of interest represents weights close to 0.5, and finally a *low* degree of interest indicates weights close to 0.

As it can be seen from table 1, the six first users (1 to 6) have *medium* or *high* degrees of interests in *motor* and *construction* topics. For them it is expected to obtain a common cluster, named cluster 1 in the table. The next six users (7 to 12) share again two topics in their preferences. They like concepts associated with *family* and *animals* topics. For them a new cluster is expected, named cluster 2. The same situation happens with the next six users (13 to 18). In this case their common preferences are the topics *beach* and *vegetation*. Their expected cluster is named cluster 3. Finally, the last two users have 'noisy' profiles, in the sense that they do not have preferences easily assigned to one of the previous clusters. However, it is comprehensible that User19 should be assigned to cluster 1 because of her high interests in *construction* topic and User20 should be assigned to cluster 2 due to her high interests in *family* topic.

Table 2. Initial concepts for each of the six considered topics

Topic	Concepts
Motor	Vehicle, Motorcycle, Bicycle, Helicopter, Boat
Construction	Construction, Fortress, Road, Street
Family	Family, Wife, Husband, Daughter , Son, Mother, Father, Sister, Brother
Animals	Animal, Dog, Cat, Bird, Dove, Eagle, Fish, Horse, Rabbit, Reptile, Snake, Turtle
Beach	Water , Sand, Sky
Vegetation	Vegetation, Tree (instance of Vegetation), Plant (instance of Vegetation), Flower (instance of Vegetation)

Table 2 shows the correspondence of concepts to topics. Note that user profiles do not necessarily include all the concepts of a topic. As mentioned before, in real world applications it is unrealistic to assume profiles are complete, since they typically include only a subset of all the actual user preferences.

We have tested our method on this simple ontology and the twenty defined user profiles. In the first step, new concepts are added to the profiles by the Constrained Spreading Activation strategy, enhancing the concept and user clustering that follows. The applied clustering strategy is a hierarchical procedure based on the Euclidean distance to measure the similarities between concepts, and the average linkage method to measure the similarities between clusters. During the execution, $N - 1$ clustering levels are computed, N being the total number of concepts. A stop criterion to set an appropriate number of clusters would be needed, but since in our case the number of expected clusters is three, the stop criterion was not necessary. Table 3 summarizes the users assigned to each final cluster and their similarities values.

It can be seen that the results are totally coincident with the expected values presented in Table 1. All the users are assigned to their corresponding clusters. Furthermore, the users' similarities values reflect their degrees of membership to each cluster. Hence the first two users of each cluster (those with high degrees of interest in their preferred topics) have the highest similarity values inside their clusters, and users 18 and 19, who had the 'noisiest' profiles, present the lowest ones. Regarding user 12, it

has to be noted that her exceedingly low similarity value is due to the low preference weights in her profile. Although Table 1 show that this user has *medium* degrees of interest for the *family* and *animals* topics, we assigned her weights close to but always below 0.5.

Table 3. User clusters and associated similarity values between users and clusters. The maximum and minimum similarity values are shown in bold and italics respectively

Cluster	Users						
1	*User1*	*User2*	*User3*	*User4*	*User5*	*User6*	*User19*
	0.522	**0.562**	0.402	0.468	0.356	0.218	*0.194*
2	*User7*	*User8*	*User9*	*User10*	*User11*	*User12*	*User20*
	0.430	**0.389**	0.374	0.257	0.367	*0.169*	*0.212*
3	*User13*	*User14*	*User15*	*User16*	*User17*	*User18*	
	0.776	**0.714**	0.463	0.437	0.527	*0.217*	

Finally, we show in Table 4 the final concepts obtained and grouped in the semantic Constrained Spreading Activation and concept clustering phases. Although most of them do not appear in the initial user profiles, they help in the construction of the clusters. Our plans for future work include studying in depth the influence of the CSA phase in realistic empirical experiments.

Table 4. Concepts assigned to the different obtained user clusters

Cluster	Concepts
1	Vehicle, Racing-Car, Tractor, Ambulance, Motorcycle, Bicycle, Helicopter, Boat, Sailing-Boat, Water-Motor, Canoe, Surf, Windsurf, Lift, Chair-Lift, Toboggan, Cable-Car, Sleigh, Snow-Cat Construction, Fortress, Garage, Road, Speedway, Racing-Circuit, Street, Wind-Tunnel, Pier, Lighthouse, Beach-Hut, Mountain-Hut, Mountain-Shelter, Mountain-Villa, Short-Oval
2	Family, Wife, Husband, Daughter , Son, Mother-In-Law, Father-In-Law, Nephew, Parent, 'Fred' (instance of Parent), Grandmother, Grandfather, Mother, Father, Sister, 'Christina' (instance of Sister), Brother, 'Peter' (instance of Brother), Cousin , Widow Animal, Vertebrates, Invertebrates, Terrestrial, Mammals, Dog, 'Tobby' (instance of Dog), Cat, Bird, Parrot, Pigeon, Dove, Parrot, Eagle, Butterfly, Fish, Horse, Rabbit, Reptile, Snake, Turtle, Tortoise, Crab
3	Water, Sand, Sky Vegetation, 'Tree' (instance of Vegetation), 'Plant' (instance of Vegetation), 'Flower' (instance of Vegetation)

4 Semantic Group Profile Modeling

As an applicative development of our method, we have experimented with building focused group profiles. After computing a multi-layered user network, and finding clusters of users with similar interests, following our previously described approach,

the profiles of such users are merged. The group profiles can be built off-line, enabling an efficient strategy for collaborative recommendation in real-time, by using the merged profiles as representatives of classes of users, whereby newcomers can be assigned to a class by comparing their profiles with the joint profile, and then be recommended items based on the group profile.

In order to combine the preferences of groups of users, a number of group modeling strategies based on social choice theory, i.e. deciding what is best for a group given the opinions of individuals, have been applied in a personalized multimedia content retrieval system. The strategies, that have been adapted to consider the semantic (weighted) preferences of our user profile representation, have been empirically tested against real subject opinions about which should be the optimal retrieved multimedia item rankings for a certain set of items and a certain group of users.

In this section, we study the feasibility of applying strategies, based on social choice theory [11], for combining multiple individual semantic profiles in our knowledge-based multimedia retrieval system. Several authors have tackled the problem combining, comparing, or merging content-item based preferences from different members of a group. Here we propose to exploit the expressive power and inference capabilities [1,12] supported by ontology-based technologies.

Combining several semantic profiles with the considered group modeling strategies we pursuit to establish how humans set an optimal multimedia items ranked list for a group, and how they measure the satisfaction of a given item list. The theoretical and empirical experiments performed will demonstrate the benefits of using semantic user preferences representations and exhibit which semantic user profiles combination strategies could be appropriate for a collaborative environment.

In [11] Judith Masthoff discusses several strategies for combining individual user models to adapt to groups. Considering a list of TV programs and a group of viewers, she investigates how humans select a sequence of items for the group to watch. Here we reproduce some of her experiments considering our personalized retrieval system and its semantic user profile representations. In this scenario, because of we have explored the combination of ontology-based user profiles, instead of rating lists, we had to slightly modify the original strategies. For instance, due to items preference weights have to belong to the range [0,1], the weights obtained for a group profile must be normalized.

The following are brief descriptions of the selected strategies.

- **Additive Utilitarian Strategy**. Preference weights from the users of the group are added, and the larger the sum the more influential the preference is for the group.
- **Multiplicative Utilitarian Strategy**. Instead of adding the preference weights, they are multiplied, and the larger the product the more influential the preference is for the group. This could be self-defeating: in a small group the opinion of each individual will have too much large impact on the product. Moreover, in our case it is advisable not to have null weights because we would lose valued preferences. Hence if this situation happens, we set the weights to very small values (e.g. 10^{-3}).
- **Borda Count**. Scores are assigned to the preferences according to their weights in a user profile: those with the lowest weight get zero scores, the next one up one point, and so on. When an individual has multiple preferences with the same weight, the averaged sum of their hypothetical scores are equally distributed to the involved preferences.

- **Copeland Rule**. Being a form of majority voting, this strategy sorts the preferences according to their *Copeland index*: the difference between the number of times a preference beats (has higher weights) the rest of the preferences and the number of times it loses to them.
- **Approval Voting**. A threshold is considered for the preferences weights: only those weights values greater or equal than the threshold value are taking into account for the profile combination. A preference receives a vote for each user profile that has its weight surpassing the establish threshold. The larger the number of votes the more influential the preference is for the group. In the experiments the threshold will be set to 0.5.
- **Least Misery Strategy**. The weight of a preference in the group profile is the minimum of its weights in the user profiles. The lower weight the less influential the preference is for the group. Thus, a group is as satisfied as its least satisfied member. Note that a minority of the group could dictate the opinion of the group: although many members like a certain item, if one member really hates it, the preferences associated to it will not appear in the group profile.
- **Most Pleasure Strategy**. It works as the Least Misery Strategy, but instead of considering for a preference the smallest weights of the users, it selects the greatest ones. The higher weight the more influential the preference is for the group.
- **Average Without Misery Strategy**. As the Additive Utilitarian Strategy, this one assigns a preference the average of the weights in the individual profiles. The difference here is that those preferences which have a weight under a certain threshold (we used 0.25) will not be considered.
- **Fairness Strategy**. The top preferences from all the users of the group are considered. We select only the $N/2$ best ones, where N is the number of preferences not assigned to the group profile yet. From them, the preference that least misery causes to the group (that from the worst alternatives that has the highest weight) is chosen for the group profile with a weight equal to 1. The process continues in the same way considering the remaining N-1, N-2, etc. preferences and uniformly diminishing to 0 the further assigned weights.
- **Plurality Voting**. This method follows the same idea of the Fairness Strategy, but instead of selecting from the $N/2$ top preferences the one that least misery causes to the group, it chooses the alternative which most votes have obtained.

Some of the above strategies, e.g. the *Multiplicative* and the *Least Misery* ones, apply penalties to those preferences that involve dislikes from few users. As mentioned before, this fact can be dangerous, as the opinion of a minority would lead the opinion of the group. If we assume users have common preferences, the effect of this disadvantage will be obviously weaker. For this reason, we shall define the individual profiles with preferences shared by the users in more or less degree.

The mechanism to apply the above strategies in the retrieval process is shown in Figure 2. Combining the semantic user profiles we shall generate a unique semantic group profile that will establish the final multimedia ranking. In the experiments we try to find the group modeling strategy that better fits the human way of selecting items when the personal and collective interests of the group have to be considered.

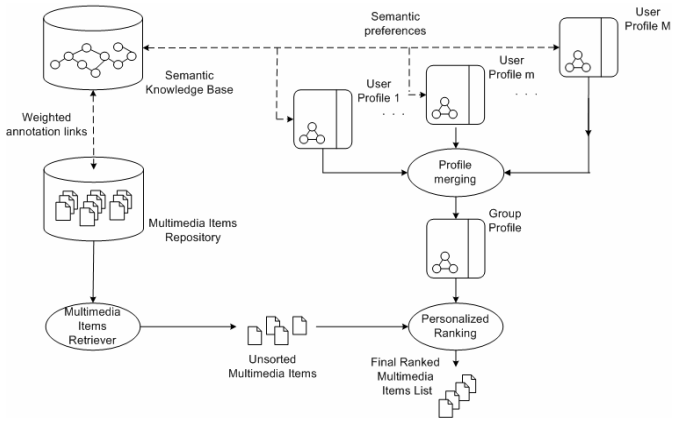


Fig. 2. User profile combination mechanism

The scenario of the experiments was the following. A set of twenty four pictures was considered. For each picture several semantic-annotations were taken, describing its topics (at least one of *beach*, *construction*, *family*, *vegetation*, and *motor*) and the degrees (real numbers in [0,1]) of appearance these topics have on the picture. Ten subjects participated in the experiments. They were Computer Science Ph.D. students of our department. They were asked to assume a group of three users with different interests. In decreasing order of preference: a) $User_1$ liked *beach*, *vegetation*, *motor*, *construction* and *family*, b) $User_2$ liked *construction*, *family*, *motor*, *vegetation* and *beach*, and c) $User_3$ liked *motor*, *construction*, *vegetation*, *family* and *beach*.

To determine which group modeling strategies give ranked lists closest to those empirically obtained from the subjects we have defined a distance that measures the existing difference between two given ranked multimedia item lists. In typical information retrieval systems, where many items are retrieved for a specific query, a user usually takes into account only the first top ranked items. In general, she will not browse the entire list of results, but stop at some top *k* in the ranking. We propose to more consider those items that appear before the *k*-th position of the strategy ranking and after the *k*-th position of the subject ranking, in order to penalize those of the top *k* items in the strategy ranked list that are not relevant for the subject.

Let Ω be the set of multimedia items stored and retrieved by the system. Let $\tau_{sub} \in [0,1]^{|\Omega|}$ be the item ranked list for a certain subject, and $\tau_{str} \in [0,1]^{|\Omega|}$ the ranked item list for a given combination strategy. We use $\tau(x)$ to refer to the position of the multimedia item $x \in \Omega$ in the ranked list τ.

Dwork et al [7] propose the Spearman distance to measure the difference between two search result lists as the average displacement of each document across the rankings. We argue that a more significant measure of impact is whether or not a result will be seen at all by the user. Since in general the user will not browse the entire list of results, but stop at some top *k* in the ranking, there are a number of documents that the user would not see (the ones ranked after the *k*-th result) in the ranking without personalization, but would see as a result of a personalized reordering, and vice versa. If we count the rate of documents in the whole collection that cross the line for each possible

value of k, and multiply it by the probability $P(k)$ that the user stops at each k, we get a loss function ranging in [0,1] that provides a measure of the effective impact (thus, the risk) of personalization in the retrieval process:

$$d(\tau_{sub}, \tau_{str}) = \sum_{k=1}^{|\Omega|} P(k) \frac{1}{k} \sum_{x \in \Omega} \left| \tau_{sub}(x) - \tau_{str}(x) \right| \cdot \chi_k(x, \tau_{sub}, \tau_{str})$$

where $P(k)$ is the probability of the user stops browsing the ranked list at position k, and

$$\chi_k(x, \tau_{sub}, \tau_{str}) = \begin{cases} 1 & \text{if } \tau_{str}(x) \leq k \text{ and } \tau_{sub}(x) > k \\ 0 & \text{otherwise} \end{cases}$$

The expression basically sums the differences between the positions of each item in the subject and strategy ranked lists, as long as they are in the top k of the strategy list and are not in the top k of the subject list. Thus, the smaller the distance the more similar the ranked lists. The problem here is how to define the probability $P(k)$. Although an approximation to the distribution function for $P(k)$ can be taken e.g. by interpolation of data from a statistical study, we simplify the model fixing $P(10) = 1$, assuming that users are only interested in those multimedia items shown in the screen at first time after a query. Our final distance is thus defined as follows:

$$D_{10}(\tau_{sub}, \tau_{str}) = \frac{1}{10} \sum_{x \in \Omega} \left| \tau_{sub}(x) - \tau_{str}(x) \right| \cdot \chi_{10}(x, \tau_{sub}, \tau_{str}) \qquad (2)$$

Observing the twenty four pictures, and taking into account the preferences of the three users belonging to the group, the ten subjects were asked to make an ordered list of the pictures. With the obtained lists we measured the distances D_{1O} with respect to the ranked lists given by the group modeling strategies. We also measure the distances against the lists obtained using semantic user profiles. Figure 3 compares the results.

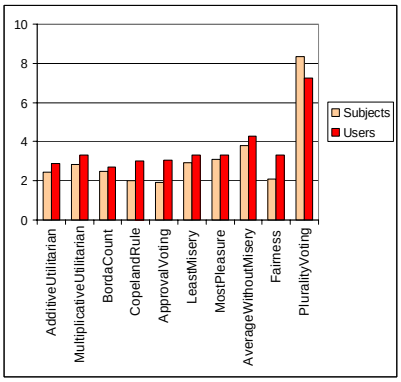


Fig. 3. Average distances D_{10} between subject lists (and user profile ranked lists) and the ranked lists obtained with the different group modeling strategies

Surprisingly, the empirical lists (those obtained from the subjects) and the theoretical (those obtained from the semantic user profiles) agree with the strategies that seem to be more adequate for modeling the group. Strategies like *Borda Count* and *Copeland Rule* give lists more similar to those manually created by the subjects, and strategies like *Average Without Misery* and *Plurality Voting* obtained the greatest distances.

5 Conclusions and future work

A variety of group-based personalization functionalities can be enabled by combining, comparing, or merging preferences from different users, where the expressive power and inference capabilities supported by ontology-based technologies act as a fundamental piece towards higher levels of abstraction.

In this work, we have presented a novel approach to the automatic identification of semantic social communities according to ontology-based user profiles. Taking into account the semantic preferences of several users, the proposed mechanism clusters the ontology concept space, obtaining common topics of interest. Each of the users is assigned to a specific cluster generating groups of users with similar interests. In a further step, these groups of users can be combined in semantic group profiles, which might be used in collaborative and recommendation systems.

Early experiments with a simple artificial problem have been done showing the feasibility of the user clustering strategy. However, more sophisticated and statistically significative experiments need to be performed in order to properly evaluate the model. Several aspects of the clustering algorithm have to be investigated using noisy user profiles: the type of clustering, the distance measure between two concepts, the distance measure between two clusters, the stop criterion that determines what number of clusters should be chosen, or the similarity measure between given clusters and user profiles. Further, a formal comparison with co-clustering methods [3,8] will have to be done.

A number of other open issues have to be addressed in future work. First of all, we plan to make more realistic experiments. In real situations, preferences can not be easily clustered. User profiles usually have noisy components and do not allow to partition the concept space in a clear way. In these cases, we hope the influence of the semantic Constrained Spreading Activation phase will be beneficial for the clustering procedure. Once the user clusters are obtained, a study of the emergent semantic social networks can be done. The preference weights of user and group profiles, the degrees of belonging of the users to each cluster and the similarity measures between clusters, constitute significant mechanisms to describe the relations between two types of social items: individuals and groups of individuals. Furthermore, the user profiles might be segmented in different preference contexts. Thus, the group modeling strategies might be improved merging certain preference contexts instead of the whole individual profiles, enriching thus the obtained semantic social networks. Finally, we are aware of the need to develop an efficient and effective automatic user profile learning algorithm. The correct concepts acquisition and their further classification and annotation in the ontology-based profiles will be crucial to the correct performance of the clustering processes.

Acknowledgements

This research was supported by the European Commission (FP6-027685 – MESH), and the Spanish Ministry of Science and Education (TIN2005-06885). The expressed content is the view of the authors but not necessarily the view of the MESH project as a whole.

References

1. Alani, H., O'Hara, K., Shadbolt, N.: *ONTOCOPI: Methods and Tools for Identifying Communities of Practice*. Intelligent Information Processing 2002, pp. 225-236, 2002.
2. Ardissono, L., Goy, A., Petrone, G., Segnan, M., Torasso, P.: *INTRIGUE: personalized recommendation of tourist attractions for desktop and handset devices*. Applied Artificial Intelligence, Special Issue on Artificial Intelligence for Cultural Heritage and Digital Li-braries 17(8-9), pp. 687-714. Taylor and Francis, 2003.
3. Cheng, Y., Church, G.M.: *Biclustering of expression data*. In Proc. of the 8th Intl. Conference on Intelligent Systems for Molecular Biology (ISM), pp. 93-103, 2000.
4. Cohen, P. R. and Kjeldsen, R.: *Information Retrieval by Constrained Spreading Activation in Semantic Networks*. Information Processing and Management, 23(2), pp. 255-268, 1987.
5. Crestani, F., Lee, P. L.: *Searching the web by constrained spreading activation*. Information Processing & Management, 36(4), pp. 585-605, 2000.
6. Duda, R.O., Hart, P., Stork, D.G.: *Pattern Classification*. John Wiley. 2001.
7. Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: *Rank Aggregation Methods for the Web*. In Proceedings of the 10th Intl. World Wide Web Conference (WWW'01), Hong Kong, 2001.
8. George, T., Merugu, S.: *A Scalable Collaborative Filtering Framework based on Co-clustering*. In Proc. of the 5th IEEE Conference on Data Mining (ICDM), pp. 625-628, 2005.
9. Liu, H., Maes, P., Davenport, G.: *Unraveling the Taste Fabric of Social Networks*. International Journal on Semantic Web and Information Systems, Vol. 2, Issue 1, pp. 42-71. 2006.
10. McCarthy, J., Anagnost, T.: *MusicFX: An arbiter of group preferences for computer supported collaborative workouts*. ACM International Conference on Computer Supported Cooperative Work (CSCW 1998). Seattle, Washington, pp. 363-372, 1998.
11. Masthoff, J.: *Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers*. User Modeling and User-Adapted Interaction, vol.14, no.1, pp.37-85, 2004.
12. Middleton, S.E., Shadbolt, N.R., Roure, D.C.D.: *Ontological user profiling in recommender systems*. ACM Transactions on Information Systems, Vol. 22(1), pp. 54-88, 2004.
13. Mika, P.: *Ontologies Are Us: A Unified Model of Social Networks and Semantics*. Proc. of the 4th International Semantic Web Conference (ISWC 2005), pp. 522-536, 2005.
14. Mika, P.: *Flink: Semantic Web technology for the extraction and analysis of social networks*. Web Semantics: Science, Services and Agents on the WWW. Vol. 3(2-3), pp. 211-223, 2005.
15. O'Conner, M., Cosley, D., Konstan, J. A., Riedl, J.: *PolyLens: A recommender system for groups of users*. 7th European Conference on Computer Supported Cooperative Work (ECSCW 2001). Bonn, Germany, 2001, pp. 199-218, 2001.
16. Vallet, D., Mylonas, P., Corella, M. A., Fuentes, J. M., Castells, P., Avrithis, Y.: *A Semantically-Enhanced Personalization Framework for Knowledge-Driven Media Services*. IADIS WWW/Internet Conference (ICWI 2005). Lisbon, Portugal, 2005.

Exploring Social Topic Networks with the Author-Topic Model

Laura Dietz

Fraunhofer Integrated Publication and Information Systems Institute (IPSI)
Dolivostr. 15, 64293 Darmstadt, Germany

Laura.Dietz@ipsi.fraunhofer.de
http://www.ipsi.fraunhofer.de

Abstract. Most approaches in Social Network Analysis have ignored the contents of collaboration artifacts. The goal of this work is to combine social network analyses with probabilistic topic models for identifying sub-communities with non-exclusive membership based on the co-authorship relation as well as topical similarity. The results of this work is integrated in the VIKE-Framework to support researchers in exploring scientific communities.

1 Introduction

One goal of Social Network Analysis is to make knowledge in a network of interrelated entities explicit. One example is the identification of communities, which are parts of the network that have strong internal and few external connections. So far, research has focused on community identification algorithms that take only the graph structure of the network, i.e. vertices and edges, as input.

In contrast to this, studies [1] indicate that for researchers the most important decision criterion for examining a person or a publication in detail is it's topic. To address this, our work provides an algorithm that puts communties of a social network in context with topics. This algorithm is embedded in a community service of the VIKE-Framework[1] [2] in order to support researchers in exploring communities of research areas relevant for their daily work.

Topics are latent concepts burried in the textual artifacts of a community. For instance "social network analysis" and "semantic web" are topics in the community this workshop addresses. It is quite intuitive, that each publication in this community addresses both of these (and other) topics, but usually puts different weights on each one. In mathematical terms, each publication has a different probability distribution over the common topics in the given community. To stress this, the term "topic mixture" is used throughout this paper.

Latent Dirichlet Allocation (LDA) [3] provides a technique for automated topic extraction from textual content, but ignores any relations among the publications that might be given as a social network. The Author-Topic Model [4]

[1] http://www.vikef.net

proceeds by incorporating contents of publications and author-of relations. In contrast to this, the Girvan-Newman Algorithm [5] succeeds in identifying communities from the co-authorship relations, but does not take the contents into account. In this work, we bring the two worlds together in a Social Topic Network (cf. Definition 1), where communities are given by relationships (e.g. co-authorship) as well as similarity in content (e.g. writing about the same topic).

Social Topic Networks can be used to provide the users of VIKEF community services with information about the main topics in communities, key persons, persons contributing to more than one community, and authors with a similar research focus. Furthermore, investigating citation relationships across communities can identify communities that are research supplyers to other communities.

In our scenario we focus on medium sized communities that have approximately 200 members, thus predefined topic taxonomies will not be considered available due to their drawbacks indicated by [6]. For that reason, we rely on unsupervised topic extraction techniques.

The next section describes the data we are working on. The third section revises the Author-Topic Model for extracting topic mixtures in an unsupervised, data-driven manner. It is followed by our approach to calculate a Community-Topic Model in Social Topic Networks. The last section includes a comparison of related work and our approach.

2 Data

Our approach builds on a collection of publications that all belong to the same (meta-)community. The authors $\boldsymbol{a_d}$ (Note, that bold characters refer to vectors, where normal characters denote entries) of each publication d are known and deduplicated (i.e. different ways of spelling the same name are resolved to map to the same person). This defines the relationship author-of$(d) = \boldsymbol{a_d}$. The contents of each publication is represented as a "bag of words", that is a vector $\boldsymbol{w_d} = (w_1, w_2, ...w_i, ..., w_{max})$ where each index i represents one vocable and w_i the number of times the vocable occurs in the document.

In the remainder of this work, we consider the entity types publication, author (also referred to as person or member) and community as depicted in Figure 1 with the relations author-of:publication→author, co-author-of:author→author and member-of:author→community.

3 Author-Topic Model

The latent Dirichlet Allocation (LDA) [3] automates extraction of common topics from a collection of texts. Instead of modelling documents $d \in \boldsymbol{d}$ independently from each other as a set of words $\boldsymbol{w_d}$, a set of latent topic variables $\boldsymbol{z}$ that are shared throughout the corpus (expressed by the random variable Z) is introduced that couple different documents via words that co-occur. Since a topic is expressed by different vocables, the number of topic variables is chosen to be much smaller then the size of the vocabulary. For each topic z a probability

distribution $\phi^z = Pr(W|z)$ over words indicates the significance of each word for the topic. A topic mixture of a publication d is expressed as a probability distribution $\theta^d = Pr(Z|d)$ over topics.

Author-Topic Model [4] is an extention to LDA that extracts topic mixtures for authors given their publications. LDA and the Author-Topic Model address the fact that an entity usually copes with more than one topic and that the topics are weighted differently. By using a dirichlet prior, they favor the association of only few topics to each entity. This way, each document and author is considered as one cohesive unit, i.e. it is more likely that words within a publication are about the same topic than words of different documents and the publications of the same author than publications of different authors.

The Author-Topic Model applies approximate inference mechanisms to calculate the model M_A with latent topic variables which approximates the given corpus best. This is achieved by maximizing the likelihood of each token given by

$$\forall d \in \boldsymbol{d}, \forall w \in d: \; Pr(w|\boldsymbol{a}_d) = \frac{1}{|\boldsymbol{a}_d|} \sum_{a \in \boldsymbol{a}_d} \phi^w \theta^a_A$$

where $\boldsymbol{\phi}$ is a distribution over significant words for each topic and $\boldsymbol{\theta_A}$ is the topic mixture for each author.

The model M_A is represended via the two sets of distributions $M_A = (\boldsymbol{\phi}, \boldsymbol{\theta_A})$. In addition, for each token j in the given document collection, that is represented by a tuple of the document $d_j \in \boldsymbol{d}$ and the word $w_j \in \{w_1, ..., w_{max}\}$ it represents, an assigment to one of the authors and topics $(w_j, d_j) \mapsto (a_j, z_j)$ is calculated. This can be used to measure the degree $\gamma^d_A = Pr(A|d)$ to which each of the authors $a \in A$ have contributed to the publication d, depending on the words and topics of other publications written by a. In addition, topic mixtures $\boldsymbol{\theta_D}$ for each publication can be calculated from the token assignments, that allow to put publications in topical relation to its authors. To stress this, we write $M_A = (\boldsymbol{\phi}, \boldsymbol{\theta_A}, \boldsymbol{\theta_D}, \boldsymbol{\gamma_A})$.

The Author-Topic Model allows to draw conclusions about the topic mixtures of entities that – although not bearing text information directly – are in some relationship to the text. Thus, the Author-Topic Model provides a way to encode influences of text on entities of the next level of abstraction, instead of drawing conclusions only on the text level as LDA does.

4 Approach to the Community-Topic Model

4.1 Algorithm for Calculation

Our algorithm for calculating a topic model for communities takes publications, its authors and the author-of relationship as inputs. It identifies communities where a person can be a member in several communities at once and calculates the degree to which this person contributes to each of those communities. Topic mixtures are extracted for each of the entity types, that is publication, person and community. The algorithm proceeds in several steps:

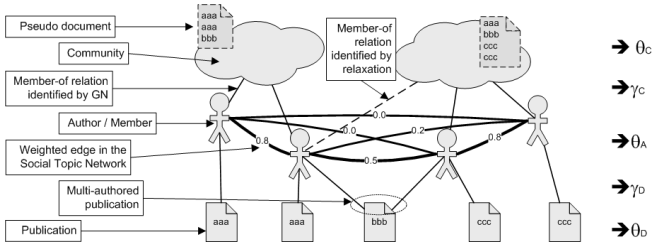


Fig. 1. Example scenario of a Community-Topic Model. Both influences, co-authorship, and writing about the same topic are reflected in the Social Topic Network. The pseudo documents for the communities are concatenations of its members' publications.

First, the Author-Topic Model is applied to the set of publications and their authors to yield the model $M_A = (\boldsymbol{\phi_A}, \boldsymbol{\theta_A}, \boldsymbol{\gamma_A})$. The Jensen-Shannon-Divergence [7] provides a measure for calculating the similarity of two topic mixtures. This similarity is used in the second step to create a graph G_{M_A} where the vertices are persons and the edges describe the similarity for each pair of authors' topic mixtures θ_A^a and $\theta_A^{a'}$. Note, that the Author-Topic Model calculates the topic mixtures θ_A from textual contents of the publications as well as the co-author-of relationship. Thus, the graph G_{M_A} can be interpreted as a weighted Social Network, where writing about the same topics is also regarded as a social relation. For that reason, we call the graph G_{M_A} a Social Topic Network.

Definition 1. *(Social Topic Network) A fully connected, undirected, weighted graph $G = (A, E, \boldsymbol{\theta_A}, w)$ is called a Social Topic Network, iff*

- *A is a set of persons,*
- *E contains edges between all pairs of persons (i.e. the graph is fully connected),*
- *$\boldsymbol{\theta_A}$ maps each vertex to its topic mixture and*
- *$w : E \to \mathbb{R}^+$ is a weight (or similarity) function for the edges, which is based on the similarity of two topic mixtures, i.e. $w(\{a_1, a_2\}) = sim(\theta_A^{a_1}, \theta_A^{a_2})$.*

The subscript in G_{M_A} indicates that the weight function w is induced by the topic mixtures of the Author-Topic Model M_A.

In the third step, we use an algorithm that identifies communities $\mathcal{C}$ in the Social Topic Network $G_{M_A} = (A, E, \boldsymbol{\theta_A}, w)$. First, we calculate communities with exclusive memberships, to yield a relation member-of: $A \to \mathcal{C}$. This can be done with the Girvan-Newman-Algorithm with the extention to weighted graphs [5,8]. Then we relax the exclusive community assignment by reinserting edges $\{a_1, a_2\}$ that cross communities, if they have nearly the weight as other edges $\{a_1, \cdot\}$ or $\{a_2, \cdot\}$ of one of the two endpoints. This heuristic allows us to identify persons that contribute to more than one community, which we assume should not be ignored.

Since this is just a heuristic, we recalculate member assignments according to a probabilistic model in the fourth step, and extract topic mixtures for the communities. We employ the Author-Topic model, but instead of applying it to publications and authors, we use communities and members as inputs. In order to do this, we generate pseudo documents d_C for each community $C \in \mathcal{C}$ by concatenating the text of publications written by authors that are identified as a member of the community in step three, i.e. $d_C := \circ\{d|\text{author-of}(d) = a \wedge \text{member-of}(a) = C, \forall a \in A\}$. We interpret the pseudo document d_C to be authored by the members of the community (cf. figure 1). This defines the relationship pseudo-author-of(d_C)$= a :\Longleftrightarrow$ member-of(C)$= a$. We now apply the Author-Topic Model to the pseudo documents d_C, the authors A, and the pseudo-author-of relationship to get the Community-Topic Model $M_C = (\boldsymbol{\phi_C}, \boldsymbol{\theta_A}, \boldsymbol{\theta_C}, \boldsymbol{\gamma_C})$.

4.2 Interpretation for a Community Service

Given the Community-Topic Model $M_C = (\boldsymbol{\phi_C}, \boldsymbol{\theta_A}, \boldsymbol{\theta_C}, \boldsymbol{\gamma_C})$ the questions raised in the first section can be answered as follows.

What are the main topics of a research community? By which authors and publications are the main topics represented? The topics z are represented by word distributions $\forall z : \phi_C^z = Pr(W|z)$. This distribution allows to generate a list of the most important words for each topic to get a first impression of what the topic is about. Human readable labels can be created manually from the list of significant words or following the approach given in [9].

What are the topics of each community? For a given (sub-)community C the topic mixture $\theta_C^C = Pr(Z|C)$ indicates which of the identified topics play an important role in this community.

Which persons are key players in a given community? Which persons are contributing to several sub-communities? The contribution of each member to a community C is modelled in the γ^C distribution over the members. The members m for which the $\gamma_C^{d_C,m}$ is very high, can be considered key players in the community. In contrast to this, persons that contribute to several communities can also be identified from γ_C.

Which persons have a similar research focus? Similarity of authors' topic mixture $sim(\theta_A^{a_1}, \theta_A^{a_2})$ can be calculated via the Jensen-Shannon Divergence as pointed out in the generation step of the Social Topic Network G_{M_A}.

Which communities are research supplyers to other communities, i.e. are cited more often than in reverse? Since we did not base the calculation of the Community-Topic Model on citation relation, we can now analyse the citations between the communities. By comparing the citation links that cross communities, we can see which communities act as research supplyers to other communities like "graph theory" is for the Semantic Network Analysis community.

5 Related Work and Discussion

Matthias Trier [10] identified the need for combining social network analysis with topics in order to provide tool support for communities of practice. For topic extraction he uses very basic techniques, since his work focusses on the requirements as well as the visualization. Our algorithm could serve as a plug-and-play extention to his work.

In [11] it is examined how topics change on a random walk through a social network of webpages. The authors identify the topic of a webpage through a categorizer. Since they are only interested in broad topics, their categorizer is trained with a given topic taxonomy from DMOZ and examples for each category. In principle, this technique may be used to investigate the topic aspects in any social network. But the main drawback is the granularity of available topic taxonomies, that are – especially for medium sized communities – often not available. This is a general problem with supervised machine learning algorithms and the reason why we rely on the unsupervised identification of common topics.

BuddyFinder-CORDER [6] is a search engine for finding experts in a community. It incorporates links structures within the community (discovered from the users' web pages) with similarity of documents that are written by community members to answer queries such as "what do your employees know about, which of your customers have they contacts with, and who works well together in teams?" [6]. The approach builds on an unsupervised text miner, that calculates the similarity of documents based on the co-occurance of named entities as well as persons' names. In general, we like the idea of comparing only named entities and ignoring words that bear less reliable semantics. Nevertheless, we suggest to re-investigate the bag-of-words approach with an alternative technique to eliminate noisy words. We assume that probabilistic latent variable models like LDA and Author-Topic Model serve this purpose, but a thorough evaluation is still to come.

Preliminary experiments have been conducted using an Author-Topic Model implementation for OpenBUGS[2] and a Java implementation of the weighted Girvan Newman algorithm [8,5]. First results seem promising although an evaluation of the usefulness requires an user study. In future, we plan to create a unified probabilistic model for detecting topical communities, topic mixtures for persons as well as documents. We assume that the unified model will lead to more precise topic mixtures for authors and documents, since inference will optimize communities and topic mixtures simultaneously.

Acknowledgements

The work described in this paper has been partly funded by the European Commission through grant to the project VIKEF under the number IST-507173.

[2] http://www.math.helsinki.fi/openbugs/

We would like to thank the other members of the VIKEF project team for the numerous discussions.

References

1. Meho, L.I., Tibbo, H.R.: Modeling the information-seeking behavior of social scientists: Ellis's study revisited. Journal of the American Society of Information Science and Technology **54**(6) (2003) 570–587
2. Niederee, C., Stewart, A., Muscogiuri, C., Hemmje, M., Risse, T.: Understanding and tailoring your scientific information environment: A context-oriented view on e-science support. Lecture Notes in Computer Science **3379** (2005) 289–298
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
4. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence, Arlington, VA, USA, AUAI Press (2004) 487–494
5. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69** (2004)
6. Zhu, J., Eisenstadt, M., Goncalves, A., Denhama, C.: Buddyfinder-corder: Leveraging social networks for matchmaking by opportunistic discovery. In Stumme, G., Hoser, B., Schmitz, C., Alani, H., eds.: Proceedings of the Workshop "Semantic Network Analysis" at ISWC 2005 Conference. (2005) 55–68
7. Lin, J., Wong, S.K.M.: A new directed divergence measure and its characterization. International Journal on General Systems **17** (1991) 73–81
8. Newman, M.E.J.: Analysis of weighted networks. Physical Review E **70** (2004)
9. Perkio, J., Buntine, W., Perttu, S.: Exploring independent trends in a topic-based search engine. In: WI'04: IEEE/WIC/ACM International Conference on Web Intelligence. (2004) 664–668
10. Trier, M.: OPUS - IT-supported Visualization and Evaluation of Virtual Knowledge Communities. Applying Social Network Intelligence Software in Knowledge Management to enable knowledge oriented People Network Management. PhD thesis, TU Berlin, Fakultaet IV - Elektrotechnik und Informatik (2005)
11. Chakrabarti, S., Joshi, M.M., Punera, K., Pennock, D.M.: The structure of broad topics on the web. In: WWW '02: Proceedings of the 11th international conference on World Wide Web, New York, NY, USA, ACM Press (2002) 251–262