

Semantic methods to capture Awareness in Business Organizations

Marcel Blattner

(Laboratory for Web Science, University of Applied Science, Switzerland
marcel.blattner@ffhs.ch)

Eldar Sultanow

(University of Potsdam, Germany
eldar.sultanow@wi.uni-potsdam.de)

Abstract: In multifarious offices, where social interaction is necessary in order to share and locate essential information, awareness becomes a concurrent process that amplifies the exigency of easy routes for personnel to be able to access this information, deferred or decentralized, in a formalized and context-sensitive way. Although the subject of awareness has immensely grown in importance, there is extensive disagreement about how this transparency can be conceptually and technically implemented. This paper introduces an awareness model in order to visualize and navigate such information in multi-tiers using semantic networks, and Web3D. To support this concept we introduce two different algorithms. The first algorithm is able to guide individuals to relevant information and topics. The second one is able to infer hidden groups (clusters) in a large company network, representing various communication channels between individuals. Both algorithms produce very promising results.

Keywords: Distributed organizations, collaboration, visualization, semantic networks, hypergraphs, spectral clustering, random walk

1 Introduction

The principle motivation for this article lies in resolving the problem of major disagreement on how to capture awareness [Gross, 1998]. Awareness is an integral CSCW (computer Supported Cooperative Work) research component, which [Dourish, 1992] defines it as follows: "...awareness is an understanding of the activities of others, which provides a context for your own activity."

There is a majority consensus on the use of semantic networks in order to portray objects including their relations to each other. A concrete implementation of semantic networks is Topic Maps (TM). A Topic Map consists of Topics, Associations and Occurrences (the so-called TAO principle). Topics illustrate things that exist in reality, which are connected to each other through Associations in their relationship. Occurrences are references to further information on Topics in external documents. The informational content is not included in the Topic Map itself.

In the course of this contribution a layered model for capturing RWA will be initially defined. To this end, collaboration data will be collected and finally evaluated during the final step, the network generation. Figure 1 shows the organization of this article.

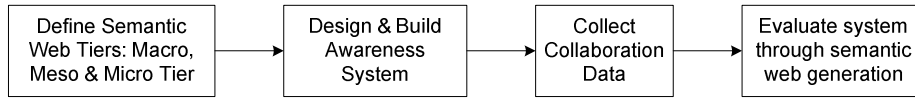


Figure 1: Procedure for implementing global Awareness with Semantic Networks

2 Three Tiers of Awareness

The model distinguishes three presentation layers, which all serve as a type of detail:

- World view (macro view): Core members of the global network and channels between them;
- Location of view (meso view): Local offices, located partners and relevant site-related infrastructure;
- View of an organization unit (micro view): workplace, roles, responsibilities and artifacts.

Personnel and activities are specifically presented to particular cases at each corresponding layer. Entities in the micro view (roles) are atomic. The elements of a layer are wired in a semantic network. An element may, in turn, be described again by elements of a semantic network in a subordinate detail [Sultanow, 2010].

2.1 Macro & Meso View

The macro view displays the locations of the core members, including their connections and available channels between these locations. It shows topics such as engineering offices and testing divisions overseas with a transfer of artifacts.

The second presented layer provides a detailed view of individual sites. It lists those sites and channels that are only noteworthy for establishments of network development at any one particular regional site. These include agencies, suppliers and depots, which are only interested in this particular regional site to implement and maintain its work. For a detailed discussion on macro and meso view along with an evaluation of requirements and benefits in business organizations, see [Sultanow, 2010].

2.2 Micro View

The third and lowest level is semantically linked to the places of employment, positions, roles and artifacts. This level details the view of individual organizational units. It further displays the principle as well as all of the available channels. Jobs are associated with artifacts (documents), whereby job descriptions or access rights act as additional information, which can be complemented in the form of occurrences.

As a topic people are assigned to their according jobs, positions and roles. Links may be between jobs, positions, given roles and separate actors. Actors have the ability to use the channels that are visually available in the macro, meso and micro views. When they are in the relevant period of use in any one given channel, then they

will be visualized. Additional information about the current activity will then be treated as an occurrence, and suitably displayed.

Activities are always addressed by at least one actor and are treated as Topics. The visualization not only show those as directly neighboring objects of the involved actors, but also at the depicted connection lines that offer the channel for this activity. Topics, and in particular, activities, may vary according to their temporal occurrence and can be faded in and out. This provides an opportunity to visualize temporal relationships.

Figure 2 shows the structure of a business unit in a Micro View. Two view-types can be made out, a static and dynamic view [Sultanow, 2010]. The static view illustrates employees, roles and artifacts. The dynamic view serves to show how actors interact with each other. This could be, for example, a phone call or a sent fax.

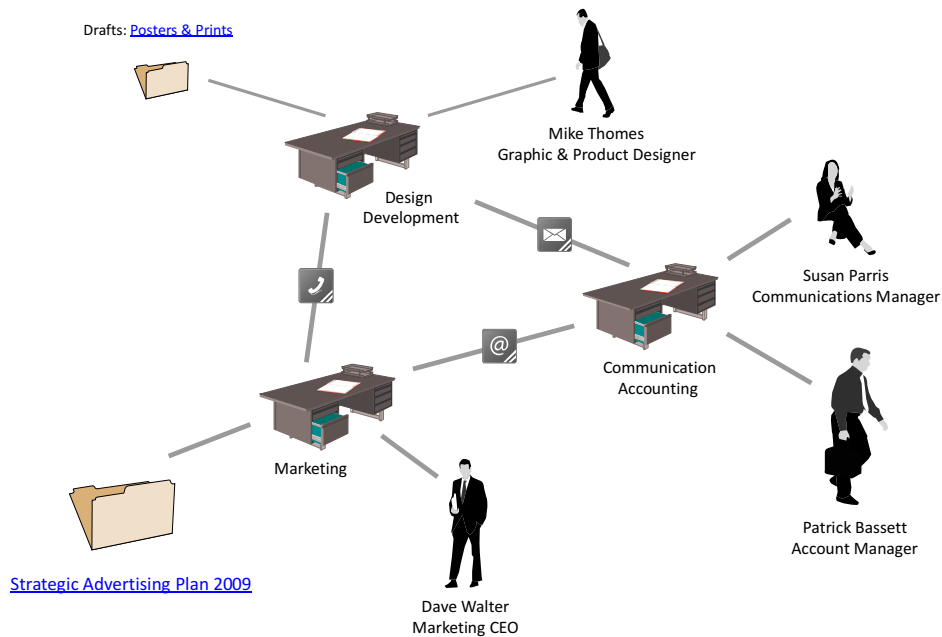


Figure 2: Semantic Network showing employees, artifacts within an institution (static view).

3 Collection of Collaboration Data

Collaboration data are very useful to infer the ‘true’ information channels in large organizations. Information, based on real communication channels is useful in optimizing processes, team building and detecting redundancies. There are various sources for such collaboration data: internal blogs, emails, Instant Messaging (IM) logs and others. Inferred data may then contain knowledge on: “who is the expert in what area” and “who is asking for what”. Methods and technologies described in this

section allow setup of a system, where organization members are able to find information and experts in an efficient way. Moreover, the presented methods allow people to cluster in groups with similar interests/questions, and to summarize similar objects/information in groups, where the similarity is intrinsically extracted from the data. This means: people can be grouped according their real communicational behavior and similar objects/information in a group is perceived as related concepts by organization members. The raw inferred information can be enriched by extracting Meta Data from structured information stored in databases or archives.

The process of data accumulation in such environments is mainly based on Data Mining and Natural Language Processing techniques.

4 System Evaluation

To illustrate a possible procedure, firstly the technical-mathematical concepts are introduced. Then two different cases are investigated:

- a) Query-Answer inference (extracting information from communications systems)
- b) Cluster inference (e.g. grouping according to role, skills and interests)

Mathematical concepts can be described as follows: Going forward, the assumption is made that relevant data has already been extracted from various channels. An efficient representation of an 'Information-Human' network is a hypergraph. A hypergraph $G(V,E)$ is a finite set V of vertices, together with a finite multiset E of hyperedges, which are arbitrary subsets of V . The incidence matrix H of a hypergraph $G(V,E)$ with $E = \{e(1),e(2),\dots,e(m)\}$ and $V = \{v(1),v(2),\dots,v(m)\}$ is the $m \times m$ matrix with $h(ij) = 1$ if $v(j)$ in $e(i)$ and 0 otherwise. Such a hypergraph can be visualized as shown in Figure 3.

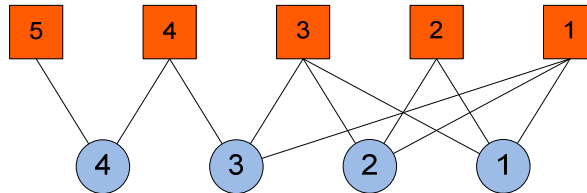


Figure 3: Hypergraph - Squares represent hyper vertices (topics) and circles represent hyper edges (people).

Red squares are the hyper vertices and represent topics or information. The blue circles represent people. In the hypergraph framework they are represented as hyperedges. Note: each edge connects more than two nodes. The interpretation of the connections is context dependent: they may represent people's special interests with regard to specific information or they may show who is the expert in what topic (information) or even who 'owns' what information. For example: node 1 (blue) is expert in topic 1, 2, 3 (red).

For instance, in ‘case a’ a relevant question related to the ‘Information-Human’ network (Figure 2) may be: if somebody is interested in information 4 and 2, what are the related topics to those? The presented technique to answer this question is based on a random walk model on hypergraphs. The calculation for an ordered ranking list, which consists of relevant concepts/information, requires a propagation matrix P . This matrix is defined as follows:

$$P(a,b) = [1 - \delta(a,b)] \frac{1}{k(a)} \sum_i [w(i)h(a,i)h(i,b)] \quad \text{with} \quad k(a) = \sum_{b,i} [w(i)h(a,i)h(i,b)]$$

as the normalization constant. $\delta(a,b)$ is the Kronecker Delta and $w(i)$ are the connection weights, $h(a,i)$ denotes the (a,i) the incidence matrix element.

Then, the algorithm consists of four steps [Blattner, 2010]:

- Forward propagation: $F(i) = P' \chi_i$
- Backward propagation: $B(i) = P \chi_i$
- Final rank: $f(i) = F(i) * B(i)$
- Sort $f(i)$ in descending order

P' is transposed from the propagation matrix P , while $*$ denotes the element-wise multiplication of two vectors and χ_i is the seed or starting information/topic (the one which is fixed to look for related concepts). To make things more clear, assume the simple case where somebody wants to infer related concepts to topic 4. Using the above algorithm and the outlined network, the final ranking list is: $f = [5, 1, 3, 2]$, (topic 4 excluded). We see that topic 5 is the most related concept. This makes sense in the above network setup, since topic 4 is only maintained by user 4 and user 3. Moreover, user 4 is elusive, only being connected to the crowd through object 4. Therefore user 4 has the strongest influence and because he is also expert in topic 5, we can expect that topic 5 is the topic that is most related to topic 4. In this simple setup we used only one seed. The algorithm reveals its real power, when choosing different seeds at the same time - mixing influences from different nodes.

To illustrate the potential of the second case, ‘case b’, we use a spectral graph theory [Chung, 1997] based approach. This technique is reported as superior compared to other methods like K-Means or ordinary PCA [Ding, 2001]. Spectral based methods have been applied in many fields like bio-informatics, recommender systems and image recognition. The clustering mechanism consists of the following steps:

- Project the hypergraph to a unipartite graph $G \rightarrow$ adjacency matrix A
- Calculate the corresponding Laplacian Matrix $L = D - A$, D is diagonal node degree Matrix
- Calculate the eigendecomposition from L
- Embed the first k non-trivial eigenvector in a k -dimensional metric (Euclidian) space
- Infer partition in this space

Here we only show how this technique is able to find network partitions (clusters), based only on the projected network structure. The projection is problematic and exceeds the scope of this paper and is omitted here.

To demonstrate the algorithm's ability to find clusters, we use a subset of the email corpus from the Enron dataset, prepared by [UC Berkeley]. Each email in the dataset is labelled with at least one of eight possible topics. In order to perform the spectral clustering method we need to calculate the adjacency matrix A . Two emails 'i' and 'j' have a weighted connection, if they are similar. A standard procedure in Natural Language Processing (NLP) has been taken to calculate similarities between each pair of emails by the use of a vector space model [Manning, 2000]. A total of 834 emails were used to group them applying the generated adjacency matrix A and the described spectral clustering algorithm. The result is shown in Figure 4. The first three non-trivial eigenvectors of the Laplacian matrix L span the metric space for the clustered emails. We observe 6-8 separated groups (clusters).

Afterwards Berkley's label-based cluster has been compared to our NLP-based cluster. Since emails have in general more than one label in Berkley's Enron dataset, we count the highest weighted label as the significant one. To be precise, it has been checked, how many emails in one of our cluster belong to the appropriate Berkley's group. Here we observed an average matching of 80% for all 834 investigated emails. In other words, the probability that an email has been classified the same as by Berkley is 80%.

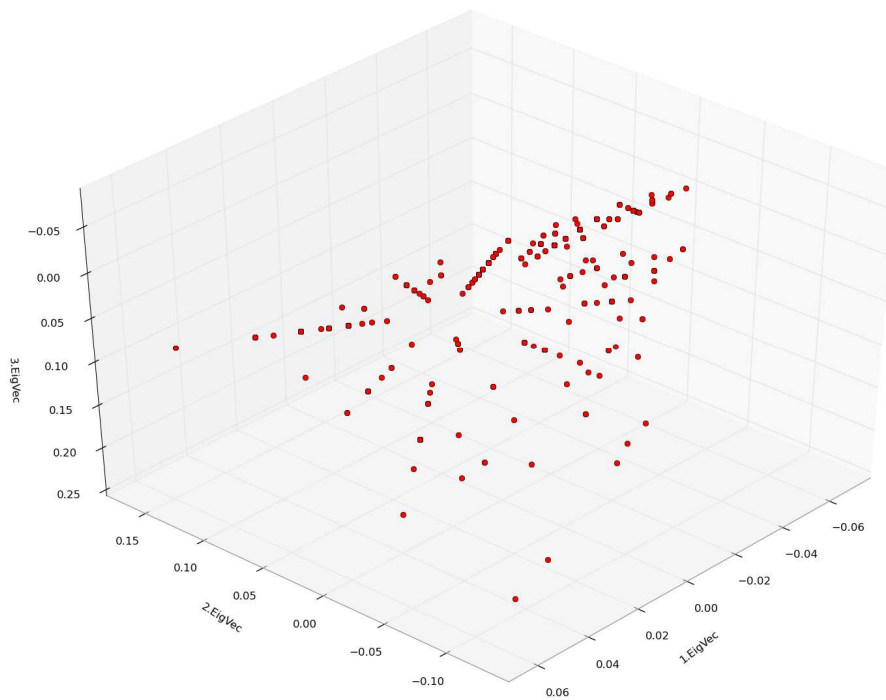


Figure 4: Clustered emails from the Enron Dataset projected in the eigenspace spanned by the first three non-trivial eigenvectors of the Laplacian matrix.

5 Conclusions

The conception in this study shows a proven method for capturing Awareness. The above methods depend on the quality of the accumulated data. That is, on the process of data accumulation as a function of space and time. The proposed methods allow visualization and optimization of processes in order to gain insight into information and knowledge flows in large organizations, one has to conduct experiments and elaborate on the data accumulation processes. This process consists of extracting information from various communication channels. The main techniques are based on Natural Language Processing and Data Mining. The methods used in this study proved powerful and efficient in the following data-capture methodologies: Query-Answer inference and Clustering detection.

References

[Blattner, 2010] Blattner, M: "B-Rank: A top N Recommendation Algorithm", IMCIC Conference on Complexity, Cybernetics and Informatics, Florida USA, 2010.

[Chung, 1997] Chung, F.: Spectral Graph Theory, American Mathematical Society Providence, RI, 1997.

[Ding, 2001] Ding, C., He, X., Gu, H., Simon, H.: A min-max cut algorithm for graph partitioning and data clustering, in: Proceedings of the first IEEE International Conference on Data Mining (ICDM), Washington, DC, USA 2001, pp. 107-114.

[Gross, 1998] Gross, T.: Von Groupware zu GroupAware: Theorie, Modelle und Systeme zur Transparenzunterstützung, German CSCW Congress: Groupware und organisatorische Innovation. September 1998.

[Dourish, 1992] Dourish, P., Bellotti, V.: Awareness and Coordination in Shared Workspaces, Proceedings of the 1992 ACM conference on Computer-supported cooperative work, ACM, 1992.

[Sultanow, 2010] Sultanow, E.; Weber, E.: Multi-Tier Based Visual Collaboration - A Model using Semantic Networks and Web3D, WEBIST 2010 - 6th International Conference on Web Information Systems and Technologies, 2010.

[UC Berkeley] http://bailando.sims.berkeley.edu/enron_email.html.