Interaction Design and Architecture(s) Journal - IxD&A, N. 5-6, 2009, pp. 35-38

A virtual discussion board: problems of computational linguistics

Cardillo Daniela
Department of Computer Science
University of Turin
corso Svizzera 185, 10149, Torino
Tel: +39 011 6706835
cardillo@di.unito.it

ABSTRACT
In this paper, the main aim is that of presenting a valid solution to a problem of computational linguistics through the presentation of a real case-study. Surveying different existing measures and techniques of semantic similarity it is proposed a valid solution to determine the degree of semantic similarity and, consequently, the relatedness of two concepts.

Categories and Subject Descriptors
I.5.3 [**Pattern Recognition**]: Clustering – *similarity measures*

General Terms
Measurement, Documentation, Reliability, Human Factors

Keywords
Social network, clustering, semantic similarity, measures of semantic distance

1. INTRODUCTION
The current situation in the web is trying to guide and accompany users through the infinite alternatives it offers. This is happening always focusing on effectiveness and efficiency of the system planning an experience always centered on the "single". The key points continue to be the transmission of the user's culture to the community of other interested users. In this way the users are tied in a culture network and each of them becomes an active supplier of concepts. This reasoning is supported through the whole paper by a deep survey of the main existing techniques of semantic similarity applied to the proposed case study.

2. WHAT IS A NETWORK?
A web of interconnected people who directly or indirectly interact with or influence the student and family. May include but is not limited to family, teachers and other school staff, friends, neighbours, community contacts, and professional support.

"A **social network** is a social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency, …."[1]

More formally, "a network contains a set of objects (in mathematical terms, nodes) and a mapping or description of relations between the objects or nodes." [10] The first and main aim of such a structure is that of creating the most suitable conditions because exists a path between two nodes having the same attributes, and a mutual relation linking one another. The chief characteristic I am going to consider in this work is that of homophily. It is defined as having one or more common social attributes; more technically, "pairs can be said to be homophilous if their characteristics match in a proportion greater than expected in the population from which they are drawn or the network of which they are a part." [14] Greater is the number of attributes in common between two nodes more likely is the chance to have an interconnection between them.

There are two thoughts about the creation of homophily considering the theory from two different points of view:

- Burt considers common rules which may bring nodes with similar attributes leading together or the reverse that is nodes with attributes in common generate common rules; [3]
- Feld and Carter, on the opposite, think that the structural location of nodes is the main cause saying that nodes may have the same attributes because the operate in the same "world" and vice versa. [7]

In general, homophily follows three fundamental rules:

1. the greater it is the more likely is the connection between two nodes;
2. it is more probable that a connection is established between people with common attributes because common nodes are promoted through common attributes;
3. the kind of connection determines a greater likelihood of a tie between nodes.

Sets of connection creates difference sectors in which the nodes can be limited giving birth to distinct areas in the region of nodes; "The region of nodes directly linked to a focal node is called the first order zone" [12] while "The nodes two steps removed from a focal node are called the second order zone, and so on." [10] In case research have to do with an entire community or very large groups it is very difficult to edge areas and apply rules to a huge amount of data so that the study is applied only to a limited sample of subjects that represent the whole network but it is more convenient to be managed. Anyway, it is convenient that the numbers of zones of the whole network is no more than three or four because all the nodes that own at a higher level zone do not have heavy effects on the focal individual or structure.

[1] http://en.wikipedia.org/wiki/Social_network

3. THE TECHNIQUE OF CLUSTERING USED IN NETWORK SEGMENTATION

Before going at the analysis of the different segments that can be created in the network, it is necessary to understand characteristics of the whole network. It is essential to usher a distinction in the establishment of a network, it can be both:

- **connected**, when a node is linked to the other, each node is reachable from another one through the path which runs from one to the other;
- **clustered**, when a node is part of an area in which nodes have mutual connections, such as relationship or affiliation to the same organization; in case the nodes are not part of the same area the parts or clusters they are part of may be confined to relatively limited *neighbourhoods* or groups. [10]

In social network clustering means a distribution of various actors of the network in a multidimensional space creating a model of it that allows observers to distinguish various components: whether the network has a limited number of actors the graphical representation helps to focus on single relationships, while in case the social network counts a huge amount of actors it gives the chance to glance at once the relationships between different groups existing inside it. The positive side of this kind of representation is that it is a natural way of representing transitivity between single nodes or whole clusters through their relationships. Moreover, this model is able to represent both direct and indirect connection between nodes and clusters taking onto account the distance between them. [9] Following in this work there will be a brief survey of clustering techniques and a deeper analysis of the (possibly) most suitable technique to create cluster at a semantic level.

4. A CASE STUDY: AN ART FORUM

4.1 Developing the system: introduction

A hypothetical case-study has been built: a web portal with a unique topic, ART & neighbourhood, this is its name. In it are advertised the major exhibitions of different typologies of art, from contemporary to naturalistic, from bohèmienne to baroque and any other kind of artistic expressions. Other than as a showcase, this portal works, at the same time, as an exchange of information about doubles and/or worst replications, photographic reproductions or photolithograph; briefly the news about the market for rich and the market for those less wealthy. On this purpose, a simple window is not enough to exchange the huge amount of information so that it has been necessary to give birth to an alternative method for the users to communicate easily and rapidly in order to give them a virtual and direct path. The first thought went to a chat; which place could be better than it to receive different kind of speeches? But soon came the requirements to keep trace of all the information in order to enrich the answers to the hypothetical questions or to personalize the offer the portal is able to do to each single user. The ideal instrument revealed to be a forum holding discussions and posting user-generated content This instrument can be exploited in different ways:

- as a simple place where talking about their own preferences and give their opinions to others, an electronic discussion group;
- as a bulletin board functioning to keep the different public sales – auction or with fixed prices – for all kinds of claims as far as the economic matter is concerned;
- as a discussion board on which to help other users becoming acquainted to "exhibitions not to be missed", a sort of list posted and enriched by single users, in which apart from the name of the exhibition and the place where it is, the user shares with others his point f view about what he saw at the exhibition. The user can describe the pieces of art, judge the whole organization of the exhibition, its arrangement including positive and/or negative judgement.

The latter point is the most important for the full use of this instrument, because from the user's point of you it can be used as a mere exchange of information, but from the webmaster's point of view it can help to create a community of people changing the forum into a social means. Cooperation between users creates a collaborative system which links strictly the single components of the community; the process of creation adopts a profile based on the model "bottom-up" so that the users do decide the topics treated. This is a strong point of value because in case the users would have been constrained in a prefixed role, it could means the end of the enrichment in knowledge, not allowing the exchange of information and the free interaction between users.

4.2 The problem of semantic similarity

In the proposed case-study it must be solved a problem of computational linguistics, that is to determine the degree of semantic similarity or, simply, the relatedness of two or more lexically expressed concepts. "Measures of similarity or relatedness are used in such applications as word sense disambiguation, determining discourse structure, text summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection, and automatic correction of word errors in text." [2] It is fundamental to remind that the semantic relatedness or distance is a different concept compared to general similarity. This latter concept is a sort of analogy or resemblance between two concepts, or even the repetition of some patterns when the concepts are compared. On the opposite, semantic relatedness is "a concept whereby a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning/semantic content"[2] adding to this notion more specific concepts as antonymy and meronymy. Various measures to evaluate the semantic distance have been used as ESA (Explicit Semantic Analysis), LSA (Latent semantic analysis), GLSA (Generalized Latent Semantic Analysis) or PMI (Pointwise Mutual Information); some of them will be analyzed in the next paragraphs and I will try to choose the most suitable/s for the case proposed.

5. A BRIEF SURVEY: TECHNIQUES IN USE

The efficient extraction of web data is often difficult, because web data does not confirm to any data standard organisation. Individuating semantic affinity among forum sentences in order to build users clusters, it may be useful to consider some issues developed in the smart web query (**SWQ**): a method for semantic retrieval of web data. The SWQ method is applied to build up a

[2] http://www.wikipedia.co.uk

search engine that facilitate the formulation of web queries. What it is interesting is the attempt to capture the semantics domain related to the user's search request in order to define the user's search needs.

This facet can be involved in the social network clustering field, for example refining keywords by exploring semantic domain. This latter should be organized in a flexible structure such as context's ontologies that define the basic terms and their relationships. This includes the vocabulary and the semantics of domain. Since terms interact with other terms originating term's relationships, directed relationships enable the SQW ontology to establish partial orderings between terms. Specifically, the "synonym" term relationship has the property of semantic distance which is not found in other term relationships. In the SQW method the semantic distance is the degree of synonymity of two terms with values ranging from 0 to 7. A score of 7 indicates very strong synonymity and 1 indicates very weak synonymity.

The objective is to build computer programs that automatically detect regularities or patterns and use these information to cluster users. Useful patterns, if found, should generalise to make accurate predictions on future data. It is also required the system provide an explicit structural description, so as to give the observer an explanation of what has been learned and an explanation of the basis for new predictions. [4] Ontologies are usually constructed by domain experts, that establish the fundamental concepts, objects, relations existing for a given community. It may be taken into account the possibility of generating ontologies automatically using hierarchical conceptual clustering, and consider certain online communities where such methods are highly appropriate, since there is no existing conceptualisation of the site resources [5]. The hierarchical conceptual clustering bases on data mining, that is, shortly, the extraction of implicit, previously unknown, and potentially useful information from data. Clustering is a data-mining task that has at its goal the unsupervised classification of a set of objects. Classification is unsupervised in the sense that there are no a-priori target classes used during training. Clustering techniques rely on the existence of some suitable similarity metric for objects [5]. For this purpose, it will be useful to employ a measure of semantic distance suitable for the ontology domain. For example the Resnik's approach is the first that brings together ontology and corpus. His measure is a formalisation of the fact that the similarity between a pair of concepts may be judged by "the extent to which they share information" [13].

Concepts may be represented probabilistically, using an algorithm like **COBWEB**, that is an incremental conceptual clustering algorithm. COBWEB is designed to produce a hierarchical classification scheme [5]. It performs a hill-climbing search - which consists of taking the current state of the search, expanding it, evaluating the children, selecting the best child for further expansion etc, and halting when no child is better than its parent – through a space of schemes, and this search is guided by an heuristic measure called category utility [8]. The category utility metric has been adopted also as a criterion for evaluating concept quality in AI systems. About it, Fisher notes that it can be viewed as a function that rewards traditional virtues held in clustering generally similarity of objects within the same class, and dissimilarity of objects in different classes [5]. COBWEB algorithm performs its search of the space of possible taxonomies and uses category utility to evaluate and select possible categorisations. It initialises the taxonomy to a single category whose features are those of the first instance. For each subsequent instance, the algorithm begins with the root category and moves through the tree. At each level it uses CU to evaluate the taxonomies resulting from [5]:

1. classifying the object with respect to an existing class;
2. creating a new class;
3. merging: combining two classes into a single class;
4. splitting: dividing a class into several classes.

In the context of an online community, one of these tasks is going to be the recommendation for the users. For example a user may request that an agent finds him/her new songs similar (or dissimilar) to those s/he has liked in the past. An ontology facilitates the fulfilment of these requirements, because similar songs will fall under the same concept, and degrees of similarity/dissimilarity will hopefully be captured in the relationships between concepts. But considering domains such as that of Smart Radio, it is not sure that an expert analysis could lead to the formalisation of concepts useful for to recommendation tasks [5]. By using such algorithms as COBWEB to cluster songs based on user ratings, it may be possible to discover structures more truly reflective of the similarities and dissimilarities between songs. The author doesn't expect that the discovered conceptual hierarchy will map onto any existing or familiar network of human concepts. Instead, the expectation is that of discovering structures that it was never feasible for human experts to detect. A further advantage is that the concept formation algorithms is incremental, in the sense that observations are not processed *en masse* [5].

6. SEMANTIC DISTANCE: EVALUATION OF DIFFERENT MEASURES

In the ART forum case-study, the objective is to detect automatically patterns contained in users messages and use these information to cluster users in order to pave the way to social interactions. Moreover, useful patterns – if found – should be generalised to make accurate predictions on future data. By the Latent Semantic Indexing (LSI), it is possible for example, to index, analyze and classify text documents. In this way it can be located similar messages near each other in this space and unrelated texts far apart of each other. LSI analyzes how terms are spread over the documents of a text corpus and creates a search space with document vectors. Moreover LSI has been developed to overcome problems with synonymy and polysemy. Since the document vectors are constructed in a very high dimensional vocabulary space, there has also been a considerable interest in low dimensional document representations. Latent Semantic Analysis (LSA) [6] is one of the best known dimensionality reduction algorithms used in information retrieval. It allows interpreting the dimensions of the resulting vector space as semantic concepts and the fact that the analysis of the semantic relatedness between terms is performed implicitly, in the flow of a matrix decomposition. It has to be noted that LSA often does not perform well on large heterogeneous collections [1].

The Generalized Latent Semantic Analysis (GLSA), instead, computes document vectors as linear combinations of term-vectors. GLSA is not based on bag-of words document vectors, but it begins with semantically motivated pair-wise term

similarities to compute a representation for terms, because terms offer a much greater flexibility in exploring similarity relations than documents. The Web offers a great resource for statistical approaches thanks to its great amount of documents [11].

In the case study it can be supposed that content bearing words, i.e. words which convey the most semantic information, will be combined into semantic classes that correspond to particular activities or relations containing synonyms and semantically related words. In this way, it seems very natural to represent terms as low dimensional vectors in the space of semantic concepts. The GLSA algorithm is formed by the following steps [11]. The authors assume to have a document collection $\boldsymbol{C}$ with vocabulary $\boldsymbol{V}$ and a large Web based corpus $\boldsymbol{W}$. It is necessary to construct the weighted term-document matrix $\boldsymbol{D}$ based on $\boldsymbol{C}$. Secondly, for the vocabulary words in **V**, it is needed to obtain a matrix of pair-wise similarities $\boldsymbol{S}$ using the large corpus $\boldsymbol{W}$. Then, combining the terms it can be obtained the matrix $\boldsymbol{U}^T$ of a low dimensional vector space representation of terms that preserves the similarities in $\boldsymbol{S}$, $\boldsymbol{U}^T \in \boldsymbol{R}^{k \times |V|}$. Finally, it occurs computing document vectors by taking linear combinations of term vectors $\boldsymbol{D} = \boldsymbol{U}^T \boldsymbol{D}$[3] The GLSA approach can combine any kind of similarity measure on the space of terms with any suitable method of dimensionality reduction. The traditional term-document matrix is used in the last step to provide the weights in the linear combination of term vectors. It has to be noted that this approach would suffer from noise introduced by infrequent and non-informative words. Finding methods of efficient filtering of the core vocabulary and keeping only content bearing words is the subject for future works [11].

7. CONCLUSIONS

This work does not claim to be a solution proposed to cluster in the best way semantic data but, rather, an overview on the most suitable measures of similarity that can be applied in the proposed case. The case-study of ART forum is, at the moment, a simple idea not yet realized but, in my opinion, it would find a wide application in the development of the social network in some web community. It would be even better to join the different communities in order to allow the subjects in them communicating each other without the needs to be part of the same net and use a unique way of clustering them. Once again this is only an idea which could be an unexplored field for future works.

8. REFERENCES

[1] Ando R. K., (2000). *Latent semantic space: iterative scaling improves precision of interdocument similarity measurement*. In Proc. of the 23rd ACM SIGIR, pages 216–223, 2000.

[2] Budanitsky A., Hirst G., (2001). *Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures*, University of Toronto, 2001.

[3] Burt R. S., (1982). *Toward a Structural Theory of Action: Network Models of Social Structure, Perception and Action.* New York: Academic Press, 1982.

[4] Chiang, R.H.L., Chua C.E.H., Storey V.C., (2001). A smart Web query for semantic retrieval of Web data, Data and Knowledge Engineering 38 (1) (2001).

[5] Clerkin P., Cunningham P., and Hayes C., (2001). Ontology discovery for the semantic web using hierarchical clustering. In *Semantic Web Mining Workshop at ECML/PKDD-2001*, Freiburg, Germany, 2001

[6] Deerwester S. C., Dumais S. T., Landauer T. K., Furnas G. W., and Harshman R. A., (1990). Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990

[7] Feld S., and Carter W. C., (1998). "Foci of Activities As Changing Contexts for Friendship", pp. 136-152 in *Placing Friendship in Context,* eds. Rebecca G. Adams and Graham Allan, Cambridge, UK: Cambridge University Press

[8] Gluck, M.A., Corter, J.E. (1985). Information, uncertainty, and the utility of categories. *Proceedings of the Seventh Annual Conference on Artificial Intelligence*, pp. 831-836, Detroit, MI: Morgan Kaufmann

[9] Handcock M. S., Raftery A. E., Tantrum J. M., (2005). *Model-Based Clustering for Social Networks*, University of Washington, 2005

[10] Kadushin C., 2004 "Introduction to Social Network Theory" available at: http://home.earthlink.net-ckadushin/Texts/, consulted on 19th may 2008.

[11] Matveeva I., Levow G., Farahat A., and Royer C., (2005). Term representation with generalized latent semantic analysis. In *Proceedings of the 2005 Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

[12] Mitchell, J. C. 1969. "The Concept and Use of Social Networks." Pp. 1-50 in *Social Networks in Urban Situations*, ed. J. C. Mitchell. Manchester, UK: University of Manchester Press

[13] Resnik, P. (1995). "Using information content to evaluate semantic similarity in a taxonomy" in *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (IJCAI-95), (cmp-lg/9511007)

[14] Verbrugge, Lois M. (1977). "The Structure of Adult Friendship Choices." *Social Forces* 56576-97

[3] The columns of $\boldsymbol{D}$ are documents in the *k*-dimensional space