Learning Vague Knowledge from Socially Generated Content in an Enterprise Framework

Panos Alexopoulos, John Pavlopoulos, Phivos Mylonas

Learning Vague Knowledge From Socially Generated Content in an Enterprise Framework

Panos Alexopoulos[1], John Pavlopoulos[2], and Phivos Mylonas[3]

[1] iSOCO, Intelligent Software Components S.A., Av. del Partenon, 16-18, 1-7, 28042, Madrid, Spain
palexopoulos@isoco.com
[2] Department of Informatics, Athens University of Economics and Business,Patission 76, GR-104 34, Athens, Greece
annis@aueb.gr
[3] National Technical University of Athens, Image, Video and Multimedia Laboratory, 9 Iroon Polytechniou St., 15773, Athens, Greece
fmylonas@image.ntua.gr

Abstract. The advent and wide proliferation of Social Web in the recent years has promoted the concept of social interaction as an important influencing factor of the way enterprises and organizations conduct business. Among the fields influenced is that of Enterprise Knowledge Management, where adoption of social computing approaches aims at increasing and maintaining at high levels the active participation of users in the organization's knowledge management activities. An important challenge towards this is the achievement of the right balance between informalities of socially generated data and the required formality of enterprise knowledge. In this context, we focus on the problem of mining vague knowledge from social content generated within an enterprise framework and we propose a learning framework based on microblogging and fuzzy ontologies.

Key words: vagueness, knowledge management, social web, microblogging, fuzzy ontologies

1 Introduction

Knowledge Management evolved over the last years to a serious management discipline that aims to enable enterprises and organizations to fully leverage their knowledge in their effort to grow more efficient and competitive. This leverage involves several key objectives such as identification, gathering and organization of existing knowledge, sharing and reusing of this knowledge for different applications and users and facilitation of new knowledge creation. Nevertheless, a dimension of enterprise knowledge that has so far been inadequately considered by the research community is that of *vagueness*.

Vagueness, typically manifested by terms and concepts like *Tall, Strong, Expert* etc., is a quite common phenomenon in human knowledge and it is related

to our inability to precisely determine the extensions of such concepts in certain domains and contexts. That is because vague concepts have typically fuzzy boundaries, that do not allow for a sharp distinction between the entities that fall within the extension of these concepts and those which do not. This is not usually a problem in individual human reasoning, but it may become one, i) when multiple people need to agree on the exact meaning of such terms and ii) when machines need to reason with them. For instance, a system could never use the statement *"This project requires many people to execute"* in order to determine the number of people actually needed for the project.

To deal with vague knowledge, a relatively new knowledge representation paradigm that has been proposed is Fuzzy Ontologies [1], extensions of classical ontologies that, based on principles of Fuzzy Set Theory [6], allow the assignment of truth degrees to vague ontological elements in an effort to quantify their vagueness. Thus, for example, whereas in a traditional ontology one would claim that ''*The project's budget is satisfactory"* or that ''*Jane is an expert at Artificial Intelligence"*, in a fuzzy ontology one would claim that *"The project's budget is satisfactory to a degree of 0.7"* and that *"Jane is an expert at Artificial Intelligence to a degree of 0.5"*.

Unfortunately, an important bottleneck in the process of developing and applying fuzzy ontologies for knowledge management is that of vague knowledge acquisition. This kind of bottleneck in traditional ontology development has been well documented in the literature and several approaches towards automating the knowledge acquisition process have been proposed [12] [10]. In fuzzy ontologies the problem is even more acute as the high level of subjectivity and context-dependence characterizing vague information makes the accurate definition of fuzzy degrees and membership functions a very difficult task. Yet only a few automatic approaches for fuzzy ontology population have so far been proposed, with the vast majority of them being based on text mining [13] [7] [3].

Contrary to above approaches, we envision the active participation of users in the vague knowledge acquisition process through a corresponding framework based on the so-called Web 2.0; a technological paradigm that facilitates and supports the active participation and collaboration of people on the Web. Our approach is inspired from works in the area of "crowdsourcing" [11] where a large group of people solves implicitly a problem or carries out a task through proper incentive mechanisms. In our case such mechanisms are required since one of the biggest bottlenecks in typical Knowledge Management systems, where end-users are supposed to actively participate, is precisely the hurdles they encounter that discourage them for keeping involved. On the other hand, Web 2.0, where users participate in an active manner, and willingly generate new content, has been adopted by companies for their internal processes within the so-called Enterprise 2.0 framework [8]. In particular, microblogging systems have been embraced as a way of fostering internal communication within the enterprise boundaries.

With that in mind, in this paper we propose a framework for automatic vague knowledge acquisition based on a semantically enhanced microblogging system

and a fuzzy ontology learning process that acts upon the social content produced by the enterprise's people within this system.

The remainder of this paper is organized as follows: in Section 2, we introduce necessary relevant background information utilized within this work. Section 3 describes the proposed vague knowledge acquisition process. Finally, we draw our conclusions and briefly describe our future work in Section 4.

2 Background and Problem Setting

2.1 Vagueness and Ontologies

Vagueness as a semantic phenomenon is typically manifested through predicates that admit borderline cases [5], i.e. cases where it is unclear whether or not the predicate applies. For example, some people are borderline tall: not clearly "tall" and not clearly "not tall". In the relevant literature two basic kinds of vagueness are identified: *degree-vagueness* and *combinatory vagueness* [5]. A predicate has degree-vagueness if the existence of borderline cases stems from the apparent lack of crisp boundaries between application and non-application of the predicate along some dimension. For example, *Bald* fails to draw any sharp boundaries along the dimension of hair quantity while *Red* can be vague along the dimensions of brightness and saturation.

On the other hand, a predicate has combinatory vagueness if there is a variety of conditions all of which have something to do with the application of the predicate, yet it is not possible to make any sharp discrimination between those combinations which are sufficient and/or necessary for application and those which are not. An example of this type is *Religion* as there are certain features that all religions share (e.g. beliefs in supernatural beings, ritual acts etc.), yet it is not clear which of these features are able to classify something as a religion.

At this point, it should be clarified that the notion vagueness is different from inexactness or uncertainty. For example, stating that someone is between 170 and 180 cm is an inexact statement but it is not vague as its limits of application are precise. Similarly, the truth of an uncertain statement, such as *"Today it might rain"*, cannot be determined due to lack of adequate information about it and not because the phenomenon of rain lacks sharp boundaries.

In an ontology the elements that can be vague are typically concepts, relations, attributes and datatypes [1]. A concept is vague if, in the given domain, context or application scenario, it admits borderline cases, namely if there are (or could be) individuals for which it is indeterminate whether they instantiate the concept. Primary candidates for being vague are concepts that denote some phase or state (e.g Adult, Child) as well as attributions, namely concepts that reflect qualitative states of entities (e.g. Red, Big, Broken etc.). Similarly, a relation is vague if there are (or could be) pairs of individuals for which it is indeterminate whether they stand in the relation. The same applies for attributes and pairs of individuals and literal values.

Finally, a vague datatype consists of a set of vague terms which may be used within the ontology as attribute values. For example, the attribute *performance*, which normally takes as values integer numbers, may also take as values terms like *very poor, poor, mediocre, good* and *excellent*. Thus vague datatypes are identified by considering the ontology's attributes and assessing whether their potential values can be expressed through vague terms.

2.2 Fuzzy Ontologies and Problem Definition

A fuzzy ontology utilizes notions from Fuzzy Set Theory in order to formally represent the vague ontological elements described in previous paragraph. The basic elements it provides include i) **Fuzzy Concepts**, namely concepts to whose instances may belong to them to certain degrees (e.g. *Goal X is an instance of StrategicGoal at a degree of 0.8*), ii) **Fuzzy Relations/Attributes**, namely relations and attributes that link concept instances to other instances or literal values to certain degrees (e.g. *John is expert at Knowledge Management at a degree of 0.5*) and iii) **Fuzzy Datatypes**, namely sets of vague terms which may be used within the ontology as attribute values (e.g. attribute *experience* mentioned above). In a fuzzy datatype each term is mapped to a fuzzy set that assigns to each of the datatype's potential exact values a fuzzy degree indicating the extent to which the exact value and the vague term express the same thing (e.g. *A consultant with 5 years of experiences is considered experienced to a degree of 0.6*)

The problem we wish to tackle can be defined as follows: Given a fuzzy enterprise ontology, what are the optimal fuzzy degrees and membership functions that should be assigned to its elements (concepts, relations and datatypes) in order to represent their vagueness as accurately as possible? In particular, given a fuzzy concept (e.g. CompanyCompetitor) and a set of its instances (e.g. a set of companies), we practically want to learn the degree to which each of these instances belongs to this concept (e.g. to what degree each company is considered a competitor). Similarly, given a fuzzy relation (e.g. isExpertAt) and a set of related through it pairs of instances (e.g. persons related to business areas), we want to learn the degree to which the relation between these pairs actually stands. Finally, given a fuzzy datatype (e.g. ProjectBudget) and the terms it consists of (e.g. *low, average, high*), we want to learn the membership functions of the fuzzy sets that best reflect the meaning of each of these terms.

3 Vague Knowledge Acquisition

As already discussed, vague pieces of knowledge are characterized by the existence of blurry boundaries and by high degree of subjectivity. As such, they are expected to provoke discussions, disagreements and debates among the enterprise's members. For example, it might be that two product managers disagree on what the most important features of a given product are or that two salesmen cannot decide what amount of sales is considered to be low. Our approach

is based on the facilitation and recording of such discussions and disagreements, through a microblogging platform, and their utilization for determining the optimal degrees and membership functions of a fuzzy ontology representing this knowledge.

In particular, the process we propose for performing vague knowledge acquisition within an enterprise consists of the following steps:

1. Identification within the enterprise of vague knowledge and conceptual modeling of it in the form of a fuzzy enterprise ontology.
2. Setting up of a microblogging platform in which the members of the enterprise are expected to participate and perform discussions and information exchange on all aspects regarding the enterprise and its environment.
3. Detection and extraction from the user generated platform's content of vague knowledge assertions, namely statements related to the elements already defined in the fuzzy enterprise ontology.
4. Calculation for each vague assertion of a strength value based on the utilization of various characteristics of the discussions they are involved in.
5. Aggregation of these assertions and automated generation of fuzzy degrees and membership functions.

In the following paragraphs we elaborate on each of the above steps.

3.1 Vague Knowledge Conceptualization

This step involves the identification of vague pieces of knowledge within the enterprise and their conceptualization in the form of fuzzy ontological elements (paragraph 2.2). Within this work we followed the implementation described in detail in the IKARUS-Onto methodology [1] which provides concrete steps and guidelines for identifying vague knowledge and conceptually modelling it by means of fuzzy ontology elements. Of course, it should be noted that, fuzzy degrees or membership functions for the above elements do not need to be defined a priori as they are expected to be automatically determined in the later stages of the process.

3.2 Microblogging Framework

Microblogging is one of the recent social phenomena of Web 2.0, being one of the key concepts that has brought Social Web to more than merely early adopters and tech savvy users. Simply put, microblogging is a light version of blogging where messages are restricted to less than a small number of characters. Yet, its simplicity and ubiquitous usage possibilities have made microblogging one of the new standards in social communication. There is already a large number of social networks and sites, with more blooming every day, that appear to have some microblogging functionalities, with Twitter [1] and Facebook[2] being the most famous.

[1] www.twitter.com
[2] www.facebook.com

Being of greater interest within the microblogging framework we examine, Twitter allows users to publish text limited to a maximum of 140 characters. On Twitter a user has to main roles, to publish tweets (writer) or to subscribe to other users and read their posts (reader). As a writer you are allowed to: 1) republish or retweet) other users posts; 2) make reference to other users within the published content (a.k.a mentions) by using the @ character before the users user name; 3) reply to another tweet, replies always start with @username (author of the tweet you are replying to); 4) include different types of resources to your post (i.e. hashtags and links); and 5) be listed by your followers. As a reader you can: 1) follow other users posts; and 2) organise into groups (lists) the users you follow.

The microblogging platform we adopt for the purposes of this work is miKrow [9], an intra-enterprise semantic microblogging tool that allows its end-users to share short messages expressing what are they doing, or more typically in a work environment, what are they working at. The platform works mostly like Twitter, with two important enhancements:

1. When users reply to a message they are able to denote the nature of their reply by using the predefined hashtags *#support* and *#attack*.
2. Users are also able to denote their agreement or disagreement to a message through a rating functionality (see figure 1)

These two features allows us to use the platform as an argumentation tool and capture the disagreements and debates over vague knowledge statements that may occur.

3.3 Detection and Extraction of Vague Knowledge Assertions

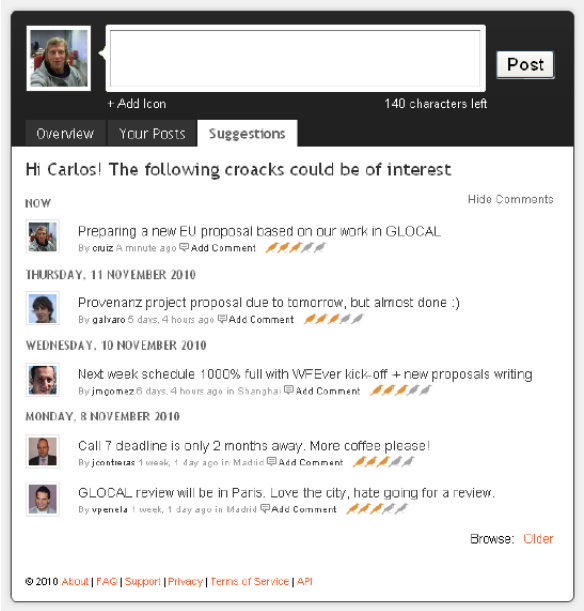
Vague knowledge assertions are practically statements related to the elements of fuzzy ontology. For example, the assertion *"The budget for the project X is low"* is related to the fuzzy datatype "ProjectBudget" while the assertion *"John knows everything about ontologies"* is related to the fuzzy relation "isExpertAt". Our goal is to detect and extract such assertions from the messages generated by the platform's users so that we can use them for determining the fuzzy degrees of their respective elements.

In order to achieve this, we use an in-house developed semantic annotation tool that, given a fuzzy ontology, is able to recognize such assertions within a piece of text. An important factor that contributes to higher levels of precision for this detection is the fact that microblogging messages are short. In any case, the detection process may be performed in a semi-automatic fashion where the correctness of the extracted assertions could be checked by the system's administrator.

3.4 Assertion Strength Assessment

To calculate the strength of the extracted vague assertions we consider their so-called "social context". The latter includes all messages that are directly or

Fig. 1. The miKrow microblogging platform



indirectly related to these assertions and may influence their validity. More formally, a social context is a tuple $G = \{U, M, A, Incl, Pub, Att, Sup, Agr, Disag\}$ where:

- U is a set of users.
- M is a set of messages.
- A is a set of vague assertions.
- $Incl$ is an assertion containment function $A \rightarrow M$ that returns for a given assertion $a \in A$ the messages it is included into.
- Pub is a message publishing function $M \rightarrow U$ that returns for a given message $m \in M$ the user that has published it.
- Att is a message attacking function $M \rightarrow M$ that returns for a given message $m \in M$ the messages that attack to it.
- Sup is a message supporting function $M \rightarrow M$ that returns for a given message $m \in M$ the messages that support it.
- Agr is a message agreeing function $M \rightarrow U$ that returns for a given message $m \in M$ the users that agree with it.
- $Disag$ is a message disagreeing function $M \rightarrow U$ that returns for a given message $m \in M$ the users that disagree with it.

Given such a context, we calculate the strength of the assertions contained in it as follows:

Let $a \in A$ be an assertion and $M_a = Incl(a)$ be the set of messages in which this assertion is contained. Then the strength of the assertion $S(a)$ is given by the average strength of these messages, namely:

$$S(a) = \frac{1}{|M_a|} \cdot \sum_{m_i \in M_a} s(m) \tag{1}$$

where $s(m)$ denotes the strength of each message and is calculated as follows:

$$s(m) = w_1 \cdot agr(m) + w_2 \cdot sup(m) + w_3 \cdot infl(Pub(m)) \tag{2}$$

In the above equation (2) w_1, w_2 and w_3 are weights denoting the relative importance, in calculating the message's strength, of the measures $agr(m), sup(m)$ and $infl(Pub(m))$ respectively. In particular, $agr(m)$ denotes the relative agreement on the message m based on the number of agreements and disagreements it has received by the users. Thus, it is calculated as follows:

$$agr(m) = \frac{|Agr(m)| - |Disag(m)|}{\sum_{m_i \in M}(|Agr(m)| + |Disag(m)|)} \tag{3}$$

Similarly, $sup(m)$ denotes the relative support to the message based on the number and strength of attacking and supporting messages. This support is recursively calculated as follows:

$$sup(m) = \frac{\sum_{m_i \in Att(m)} s(m_i)}{|Att(m)|} - \frac{\sum_{m_j \in Sup(m)} s(m_j)}{|Sup(m)|} \tag{4}$$

Finally, $infl(Pub(m))$ denotes the overall influence of the user who has published the message (and thus has made the assertion). This influence is generally relevant to the number of users that follow the message publisher but also to the persons expertise on the messages topic. For that, we derive the exact influence score using Topic-Sensitive PageRank Algorithm [4] which was originally proposed for ranking web pages according to their relative importance, given a set of representative topics. In our case, in the place of pages we have users and the topics are in practice the business areas in which these users are interested or considered expert within the enterprise.

3.5 Generation of Membership Functions and Fuzzy Degrees

A fuzzy ontology may be considered as a tuple $O_F = \{C, R, I, T, i_C, i_R, D\}$, where

- C is a set of fuzzy concepts.
- I is a set of instances.
- R is a set of fuzzy binary relations that may link pairs of concept instances.
- i_C is a fuzzy concept instantiation function $C \times I \rightarrow [0, 1]$.

– i_R is a fuzzy relation instantiation function $R \times I \rightarrow [0, 1]$.
– D is a set of fuzzy datatypes. Each $d \in D$ is itself a tuple $\{T, X, m\}$ where T is the set of linguistic terms of the datatype that refer to a base variable whose values range over a universal set X and m is a function that, for each linguistic term $t \in T$, relates the values of X to a fuzzy degree.

Based on this formalization, our goal is practically to learn the functions i_C and i_R and m. To do that we utilize the extracted assertions along with their calculated strengths. In particular, given an instance $i \in I$ and a concept $c \in C$, the related assertions form a set $A_{i,c} = \{s_1, s_2, ..., s_n\}$ where s_j is the strength of the j^{th} assertion. Based on this set we want to determine a single fuzzy degree d for the pair $\{i, c\}$. Similarly, for two given instances $i_1, i_2 \in I$ and a relation $r \in R$, the related assertions form a similar set $A_{i_1,i_2,r} = \{s_1, s_2, ..., s_n\}$. Again our goal is to determine a single fuzzy degree d for the triple $\{i_1, i_2, r\}$.

For example, if $i_1 = John$, $r = isExpertAt$ and $i_2 = MachineLearning$ and we have managed to extract a number of relevant assertions, each with some strength, we want to aggregate these strengths into a single degree that denotes how expert is actually John at Machine Learning. To do that we first compute the mean value of all the strength values observed in the assertion set. Then we estimate confidence intervals and we only allow those mean values with significance level of no less than 0.05. In most cases we expect to have most assertions gathered very close to a single mean value which can then be considered as the degree of the relevant ontological statement. In case many assertions seem to be out of the confidence intervals, then that's an indication that the statement's interpretation might be context-dependent. In such a case, we may isolate these contexts by performing clustering on the assertion set.

On the other hand, for a given term t of a fuzzy datatype $d \in D$, the related assertions form a set $A_{t,d} = \{(v_1, s_1), (v_2, s_2), ..., (v_n, s_n)\}$ where v_j is the actual value the term t refers to in the j^{th} assertion (for example, in the assertion *"The budget for the project X is low"* v is the actual value of project X). Our goal in this case is, based on the pairs of these values and strengths, to determine the optimal function fuzzy membership function m that links them. This is a well-studied problem in the area of fuzzy expert systems and several related methods that construct such functions from training data have been proposed [2]. Therefore our approach involves reusing such methods in our framework and through extensive experimentation determining the optimal one for the kind of training data our microblogging platform produces.

4 Conclusions and Future Work

In this paper we proposed a framework for automatic vague knowledge acquisition in enterprise settings, based on a semantically enhanced microblogging system and a fuzzy ontology learning process that acts upon the social content produced by the enterprise's people. The key characteristic of our approach is the utilization of the content's social features, like the relative agreement and

support that microposts enjoy or the status and influence of the users, in order to assign strengths to vague assertions.

In the future we intend to apply our framework in an actual enterprise setting and evaluate its effectiveness in acquiring vague knowledge. This evaluation will focus on two dimensions: i) the ability of the microblogging approach in producing rich social context over the vague knowledge and ii) the accuracy of the fuzzy ontology degrees and membership functions learned using this context.

References

1. Alexopoulos, P., Wallace, M., Kafentzis, K., Askounis, D.: IKARUS-Onto - A Methodology for Developing Fuzzy Ontologies. Knowledge and Information Systems, pp. 1-29, Springer (2011)
2. Chen, S.M., Tsai, F.M.: A New Method to Construct Membership Functions and Generate Fuzzy Rules from Training Instances. Information and Management Sciences 16(2), pp. 47-72 (2005)
3. Chen, W., Yang, Q., Zhu, L., Wen, B.: Research on Automatic Fuzzy Ontology Generation from Fuzzy Context. In: Proceedings of the 2009 Second International Conference on Intelligent Computation Technology and Automation,, Vol. 2. IEEE Computer Society, Washington, DC, USA, 764-767 (2009)
4. Haveliwala, T.H.: Topic-sensitive PageRank. In Proceedings of the Eleventh International World Wide Web Conference, Honolulu, Hawaii (2002)
5. Hyde, D.: Vagueness, Logic and Ontology. Ashgate New Critical Thinking in Philosophy (2008)
6. Klir, G., Yuan, B.: Fuzzy Sets and Fuzzy Logic, Theory and Applications. Prentice Hall (1995)
7. Lau, R.Y., Song, D., Li, Y., Cheung, T.C.H., Hao, J.X.: Toward a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning. IEEE Transactions on Knowledge and Data Engineering, 21(6), 800-813 (2009)
8. McAfee, A.: Enterprise 2.0: The dawn of emergent collaboration. MIT Sloan Management Review, 47(3) (2006)
9. Penela, V., lvaro, G., Ruiz, C., Cordoba, C., Carbone, F., Castagnone, M., Gomez-Perez, J.M., Contreras, J.: miKrow: Semantic intra-enterprise micro-knowledge management system. In Extended Semantic Web Conference (2011)
10. Reichartz, F, Korte, H, Paass, G.: Semantic relation extraction with kernels over typed dependency trees. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10). ACM, New York, NY, USA, 773-782 (2010)
11. Surowiecki, J., Silverman, M.: The wisdom of crowds. American Journal of Physics 75, 190 (2007)
12. Szumlanski, S, Gomez, F.: Automatically Acquiring a Semantic Network of Related Concepts. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM-10). pp. 1928. Toronto, Ontario (2010)
13. Zhang, F., Ma, Z.M., Fan, G., Wang, X.: Automatic fuzzy semantic web ontology learning from fuzzy object-oriented database model. In Proceedings of the 21st international conference on Database and expert systems applications: Part I (DEXA'10), Pablo Garcia Bringas, Abdelkader Hameurlain, and Gerald Quirchmayr (Eds.). Springer-Verlag, Berlin, Heidelberg, 16-30 (2010)