Ingeniare. Revista Chilena de Ingeniería
ISSN: 0718-3291
facing@uta.cl
Universidad de Tarapacá
Chile

Beck-Fernández, Héctor; Nettleton, David F.
Identification and extraction of memes represented as semantic networks from free text online forums
Ingeniare. Revista Chilena de Ingeniería, vol. 23, núm. 1, enero, 2015, pp. 50-58
Universidad de Tarapacá
Arica, Chile

Identification and extraction of memes represented as semantic networks from free text online forums

Identificación y extracción de memes representados como redes semánticas desde foros en texto libre

Héctor Beck-Fernández[1] David F. Nettleton[2]

Recibido 17 de abril de 2014, aceptado 11 de agosto de 2014
Received: April 17, 2014 Accepted: August 11, 2014

ABSTRACT

Memes have recently come into vogue in the context of 'viral' transmission of basic information units in online social networks. However, from their original general definition in a sociological context, there is still much work to be done from an information technology viewpoint. This includes such issues as how to process memes from real text corpus, formal definitions for knowledge representation, meme refinement and selection. In order to address these issues, in this paper we adapt definitions from the semantic network and information retrieval fields to extract memes as semantic networks from free text documents, and then we present some examples in the context of a simple online forum.

Keywords: Memes, semantic networks, information retrieval, free format text.

RESUMEN

Recientemente los memes han estado en boga en el contexto de la transmisión "viral" de unidades básicas de información en las redes sociales online. Sin embargo, a partir de su definición general original en un contexto sociológico, todavía hay mucho trabajo por hacer desde el punto de vista de la informática. Esto incluye cuestiones como la forma de procesar los memes de corpus textual, las definiciones formales de la representación del conocimiento, el refinamiento de meme y su selección. Con el fin de abordar estas cuestiones, en el presente trabajo se adaptan definiciones desde los campos de redes semánticas y de recuperación de información para extraer los memes como redes semánticas de documentos de texto libre, y luego entregamos algunos ejemplos en el contexto de un foro online simple.

Palabras clave: Memes, redes semánticas, recuperación de información, texto en formato libre.

INTRODUCTION

As defined in a sociological context by Dawkins [1] and Blackmore [2], a meme is understood as a basic element of useful knowledge, or meta-information, which can be transmitted from one individual to another. However, from an information technology point of view, many technical and implementation challenges remain, such as how to identify and extract key memes from free text document corpuses. The study of memes has a high potential utility for understanding and modelling information diffusion/influence in Online Social Networks and applications such as recommender systems [3]. Hence the work is motivated by the technical challenges on the one hand, and the potential of the application of the results, on the other.

[1] Área de Ingeniería en Computación e Informática. Universidad de Tarapacá. 18 de Septiembre 2222. Arica, Chile. E-mail: hbeck@uta.cl
[2] Department Information Technology and Communications. Universitat Pompeu Fabra. Tànger, 122-140, 08018. Barcelona, España. E-mail: david.nettleton@upf.edu

In this paper the main focus will be on the problem of defining and identifying semantic network type structures in free text, which can then be used to represent memes. We use information retrieval and semantic networks concepts to identify and extract the key memes from a larger candidate set. A simple example is given of an online forum of comment posts to illustrate how the framework could be applied in practise.

The structure of the remainder of the paper is as follows: in the second section we present the state of the art and related work; in the third section we present the definitions for the documents, semantic network concepts (entities and relations) and memes; in the fourth section we given some examples for the definitions of section three; in the fifth section we consider meme metrics and how we can use them to identify the 'top' memes extracted using the definitions and examples of sections three and four; in the final section we give the conclusions.

STATE OF THE ART AND RELATED WORK

The term "meme" was originally defined by Dawkins in [1], and has been recently applied to the study of how information spreads through Internet and Online Social Networks (OSNs).

According to Dawkins [1] a "meme" or a "memetype" is similar to a "gene" or "virus". It consists of a basic unit of information circulating among a community, and research from a social sciences perspective has studied how it serves as a mechanism to propagate cultural and social evolution [4]. In [4], Heylighen and Chielens compare the 'meme' with the 'gene' and formalize the following memes properties: 'longevity', the duration that an individual meme survives; 'fecundity', the reproductive activity of a meme; 'copy-fidelity', the degree to which a meme is accurately reproduced.

In [5], Bordogna and Pasi propose a schematic definition for memes using an OWL schema, followed by the definition of several operators to extract memes from online blog posts using information retrieval methods and n-grams (contiguous sequences of n items from a given sequence of text). A fuzzy-type matching is performed to evaluate the fidelity of a given blog post to an original meme description. Finally, the longevity is considered by ordering the text entries by their timestamp and taking into consideration the fidelity.

Leskovec, Backstrom and Kleinberg in [6] developed a framework for tracking short textual memes in an online news media environment, identifying a broad class of memes that exhibit a wide spread and rich variation on a daily basis. Simmons, Adamic and Adar in [7] presented a study about meme mutation in social networks. They uncovered patterns in the rate of appearance of new variants, their length and popularity, and developed a simple model that is able to represents these attributes. Nettleton in [8] presents a wide-ranging survey of OSN analysis, covering themes such as 'influence and recommendation' and 'information diffusion', which includes contextual entity tracking using memes. Baydin and López de Mántaras [9] present an evolutionary algorithm based on the concept of memes. They used semantic networks to represent the individual pieces of information, and employed the 'genetic' concepts of crossover and mutation to model changes over time. Their method was tested on synthetically generated examples.

Now we will briefly summarize some of the literature with respect to the extraction of semantic networks from text. Szumlanski and Gomez in [10] extracted semantic networks based on frequency and concept affinity from Wikipedia texts using the WordNet [11] ontology database to identify related concepts. In [12], Jiang and Conrath describe a semantic similarity metric based on corpus statistics and a lexical taxonomy. They present an approach for measuring semantic similarity/distance between words and concepts which uses a distributional analysis of the corpus data. In [13], Chen, Gangopadhyay, Karabatis, McGuire and Welty deals with the elicitation of semantic networks based on concepts relevant to the data mining of specific datasets. In [14], Kok and Domingos, present an unsupervised approach to extracting semantic networks from large volumes of text. They use the TextRunner system [15] to extract tuples from text, and then induce general concepts and relations from them by jointly clustering the objects and relational strings in the tuples. Their approach is defined in Markov logic using four basic rules to extract meaningful semantic networks.

EXTRACTION OF SEMANTIC NETWORKS AS MEMES FROM FREE FORMAT TEXT

In this Section we present the definitions for the meme environment (documents, concepts and relations) which will allow us to identify the key memes, represented as semantic networks, in a free format document corpus.

Introduction

In the following we will define the two main data processing steps (processes 1 and 2) and their corresponding definitions (1 to 6).

Process 1: This process acts on the complete document set *D* to identify the key concepts and relations. It is comprised of *Definitions 1* to *3*. The objective of *Definition 1* is to identify the most relevant subset of documents and key concepts from the complete document corpus. Then, *Definitions 2* and *3* identify the relationships between the key concepts. We note that *Definitions 1* to *3* act on the complete document corpus *D*.

Process 2: This process acts on individual documents to compact the semantic networks (eliminate redundant relations and identify the minimal semantic networks). It is comprised of *Definitions 4, 5* and *6*, which deal with eliminating redundant relations and finding the minimum semantic networks between concepts. We observe that *Definitions 4, 5* and *6* act on individual documents *d*.

We note the importance of the use of *thresholds* in the processing. The thresholds are determined statistically from the probability distributions of the corresponding metrics. The threshold can be defined by a quartile limit or by point of inflexion, from the corresponding distribution.

Definitions which define the extraction of semantic network as memes

Definition 1. A ***concept*** is an *n*-gram[3] (excluding *stopwords*[4], in the information retrieval sense) that is present in a significant number of documents in a document collection. Formally, let *D* be the total document collection. Then, an *n*-gram x_i is a ***concept*** when it satisfies the condition:

$$p_D(x_i) = \frac{\text{Nº of documents in } D \text{ which contain } x_i}{\text{Nº of documents in } D} = \frac{|D(x_i)|}{|D|} > \alpha \quad (1)$$

where $\alpha \in [0,1]$ a is a value known as threshold which is user defined. The threshold α indicates the percentage of documents containing an *n*-gram to be considered a concept. How is this value chosen? Low values of α will obtain many concepts; on the other hand, higher values of α will obtain fewer concepts. As we consider a document collection which is a free text comments forum, we are interested in those concepts that have most presence in the discussion. As an initial approximation, we could choose a moderately high value for α, in the order of 0.70 ± 0.05. Empirically, we could consider the three highest deciles in a frequency distribution table of candidates for concepts. Other definitions can be found in [16-17].

In a given free text block written by a user of an online community, some concepts will be related to each other, in a way that has meaning for that community. Concepts such as "democracy", "is" and "participation" have little meaning when each of them is taken in isolation, however if they are related by means of a verbal expression (which may be another concept), then they acquire much more meaning. For example, "participation is democracy". In this context we must determine which concepts are co-occurrences and which are related.

We recall that two concepts are a *co-occurrence* if they are at a distance of less than *n* words apart, in the same sentence, or in the same paragraph. In the first case, a limit of 4 can be placed on the value of *n*; an interesting study on the co-occurrence of words is to be found in Ferrer-i-Cancho and Solé's paper [18]. The following Definition 2 provides a way to determine related concepts.

Definition 2. ***Relationship related*** (*R*): Let x_i and x_j be concepts in a document collection *D*, $x_i R x_j \Leftrightarrow p_D(x_i \mid x_j) > \gamma \wedge p_D(x_j \mid x_i) > \gamma$; where:

$$p_D(x_i \mid x_j) = \frac{p_D(x_j, x_i)}{p_D(x_j)}, p_D(x_j, x_i) = \frac{|D(x_i, x_j)|}{|D|} \quad (2)$$

3 Word or group of consecutive words where *n* is the number of words making up the sequence.

4 Examples of lists of *stopwords* can be found at http://www.ranks.nl/resources/stopwords.html

and $|D(x_i, x_j)|$ represents the number of documents in D in which x_i, x_j are found as co-ocurrences.

Thus, *Definition 2* will be true if both concepts appear together in many documents in the collection D. In this case, the threshold $\gamma \in [0,1]$ indicates a measure to determine when both concepts are considered related, and is assigned by the user and verified empirically. Again, high values of γ could provide few concepts and low values of γ could provide many related concepts. By extracting the concepts that correspond to verbs, nouns and adjectives, a syntactic and semantic representation of text can be obtained in the form of semantic networks.

A semantic network (SN) is a notation that allows us to represent ideas with meaning and which represent knowledge. An SN is represented by a graph in which the nodes are concepts (nouns and adjectives) and the arcs are the relationships (verbal expressions) between them [19]. Figure 1 shows a semantic network with 9 concepts and 3 distinct relationships. This form of notation is used, for example, in the fields of natural language processing and information retrieval [20], among others.

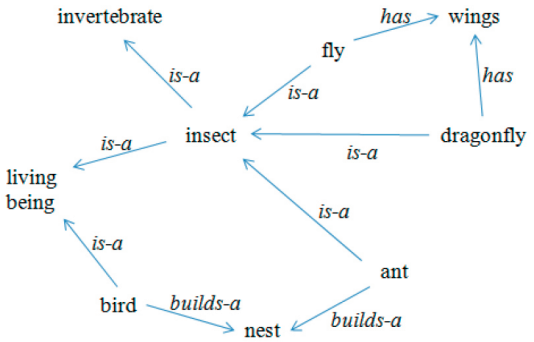

Figure 1. Semantic network with 9 concepts and 3 distinct relationships (*has*, *is-a*, *builds-a*).

Many kinds of relationships can be derived from a semantic network; we will only consider those which are relevant to our present work. As a starting point, we will use some of the definitions given in [17] by Oh, Kin, Park and Yu, changing the notation and adapting them to the present context.

Definition 3. ***Relationship superset-subset***. If the document set of $x_i (D(x_i))$ is included in the document set of $x_j (D(x_j))$ then we say that $D(x_j)$ is a *superset* of $D(x_i)$ and we denote $D(x_j) \rightarrow D(x_i)$, formally:

$$D(x_j) \rightarrow D(x_i) \Leftrightarrow p_D(x_j \mid x_i) > \delta \text{ and } p_D(x_i \mid x_j) < \delta \quad (3)$$

In this case, $\delta \in (0,1)$ can be calculated empirically using the equation:

$$|D(x_i)| * \delta < |D(x_i, x_j)| < |D(x_j)| * \delta \quad (4)$$

where $|D(x)|$ is the cardinality of set $D(x)$.

From definitions 1, 2 and 3 we have now identified the concepts as the most important terms in a document set. If we consider that x_i is a candidate concept and c_i is a chosen concept whose frequency in the document set is above the given thresholds, then $x_i \rightarrow c_i$, when the given thresholds α, γ and δ are satisfied. In the following definitions 4 to 6 we will consider a given document d_q belonging to document set D.

Definition 4. ***Relationship redundant***. A relationship $d_q(x_i) \rightarrow d_q(x_j)$ is *redundant* if there exist one o more concepts such that $d_q(x_i) \rightarrow d_q(x_k) \rightarrow \ldots \rightarrow d_q(x_j)$ in a semantic network.

Definition 5. ***Closest Superset***. Let $C_x = \{d_q(c_1), d_q(c_2), \ldots, d_q(c_k)\}$ the set of *Supersets* of x, i.e. C_x is the set of all $d_q(x_i)$ such that $d_q(x) \rightarrow d_q(c_1)$, $d_q(x) \rightarrow d_q(c_2)$, …, $d_q(x) \rightarrow d_q(c_k)$. The *Closest Superset* of x is the smallest of all $d_q(c_i)$.

Definition 6. ***Minimal Semantic Network***. Let graph $\mathcal{G} = (\mathcal{C}, \mathcal{R})$ be a semantic network where $\mathcal{C}$ is a set of concepts and $\mathcal{R}$ is the set of relationships between concepts. A semantic network $\mathcal{G}' = (\mathcal{C}, \mathcal{R}')$ with $\mathcal{R}' \subset \mathcal{R}$, is a *minimal semantic network* if, for all relationships $r_k = (c_i, c_j, Type) \in \mathcal{R}'$, $d_q(c_i)$ is the *closest superset* of $d_q(c_j)$, and where *'Type'* is the set of possible relations.

With respect to definition 6, we note that each relationship in a semantic network can be expressed by a triple (x_i, x_j, *type*) where x_i and x_i are concepts and "*type*" is the type of relationship between x_i and x_j. In [17] Oh, Kin, Park and Yu, proved that in a minimal semantic network the relationships between concepts are not redundant.

EXAMPLES

In this Section, with reference to Figures 2, 3 and 4, we will give an example of each of the aspects we have described in Section 3. We note that the objective of the future work will be to automate the process as much as possible, however we envisage a semi-automatic scheme which may require some manual annotation of the original text and semi-supervised processing in other steps, such as in [18]. Although these implementation details are out of the scope of the current paper, we can say that in order to construct the semantic networks, we would need to distinguish between the entities and the relations from the initial set of concepts. This could be done using natural language processing software tools and a relationship-instance repository together with WordNet[11], *http://wordnet.princeton.edu/*, in order to identify entities (e.g. nouns, adjectives) and relations (e.g. verbs, adverbs).

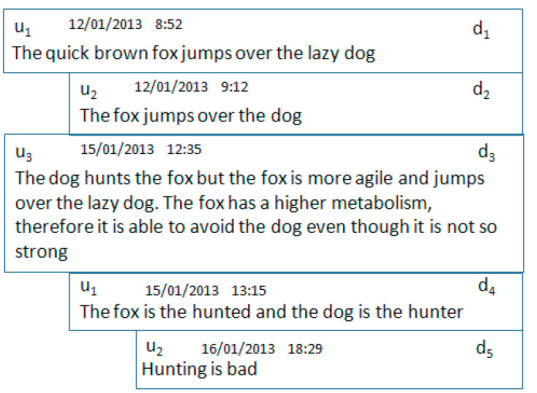

Figure 2. Online forum example: user's posts, with date and timestamps.

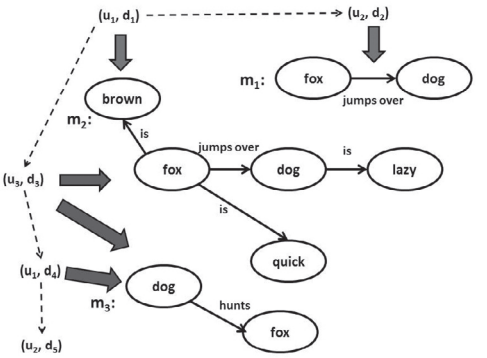

Figure 3. Online forum example: documents (comment texts), users and memes.

Firstly, in Figure 2 we see a simplified example of a typical online comments forum for a newspaper article. That is, a newspaper publishes an article about a given theme and below the article the registered users are allowed to post their opinions. What typically happens is that users with differing opinions create a debate in which some users state their opinions and other users either support or reject all or part of those opinions.

We observe in Figure 2, that user 1 has posted a comment, which is replied by user 2. Then user 3 posts a new comment, which is replied by user 1, whose comment is in turn replied by user 2. We can clearly see that the central concepts are about foxes and dogs

Concepts: correspond to the search terms, which can be entities and/or relations. In Figure 3, the semantic networks formed include the entity concepts 'fox', 'dog', 'brown', 'quick', 'lazy', and the relation concepts 'is', 'jumps-over', 'hunts'. As mentioned, semi-automatic tools exist for identifying syntax structures, however we must not underestimate the difficulty of correctly identifying the relations between entities, especially when a concept has different meanings dependent on the context. For example, the concept 'quick' can be a noun, adjective or adverb. For the present work, we assume a manual revision of the ambiguous cases. In Table 1 we see the concepts, their syntactic classification and the corresponding assignment as entity or relation.

Table 1. Concepts, syntactic categories and assignments as entity or relation.

Concept	Possible syntactic categories	Chosen syntactic category	Entity or relation
fox	noun, verb	noun	entity
dog	noun, verb	noun	entity
brown	noun, verb, adjective	noun, adjective	entity
quick	noun, adjective, adverb	adjective	entity
lazy	adjective	adjective	entity
hunt	verb, noun	verb	relation
jumps over	noun, verb adjective, adverb	verb, adverb	relation
is	noun, verb	verb	relation

Documents: a document is a block of text (comment) written by a user. In information retrieval, if we formulate a query to search for a set of terms (or concepts), such as {fox, dog}, the query will return a set of documents in which one or more (depending if the query is AND or OR) of the query terms appears. Hence, a document will contain one or more concepts which are susceptible to be formed into one or more semantic networks. In Figures 2 and 3 we see there are five documents, designated as d_1 to d_5.

Semantic network (candidate meme): a semantic network is made up of two or more entity concepts which are related by one or more relation concepts. A document may contain one or more semantic networks, made up of corresponding concepts. In Figure 3 we see that we have extracted three significant memes from all the potential semantic networks which can be constructed from the respective texts. Later, in the Section 'Incorporation of the Meme Metrics', we will consider how we can use the meme metrics to identify the most significant memes.

Superset-subset: If a set of documents Sd_1 is included within another set of documents Sd_2 then Sd_1 is a subset of Sd_2 and Sd_2 is a superset of Sd_1. This is related to the information retrieval concept of document retrieval sets corresponding to queries made of one or more query terms. In the current context the query terms would be the concepts making up the memes, that is, each meme is a potential query. With reference to Figure 3, consider the following example: the query {fox, dog, jumps} retrieves the set of documents $Sd_1 = \{d_1, d_2, d_3\}$; the query {fox, dog} retrieves the set of documents $Sd_2 = \{d_1, d_2, d_3, d_4\}$, which is a superset of document set Sd_1. Likewise, Sd_1 is a subset of Sd_2.

Redundancy: a relation (link) between two concepts is redundant if it already exists via another path. With reference to Figure 4a, we see that the link between 'fox' and 'wolf' is redundant because it is already implicit (inherited) through the links between 'fox', 'dog' and 'wolf'.

Closest superset: the smallest superset with respect to a given subset. Returning to the example of Figure 3, consider three queries, those we defined previously, Sd_1 and Sd_2, and a new one $Sd_3 = \{$fox, lazy, dog$\}$ which returns documents $\{d_1, d_3\}$. Hence, the smallest superset with respect to Sd_3 will be Sd_1, as opposed to Sd_2, given that Sd_2 contains four documents whereas Sd_1 contains only three. In Table 2 we see the queries and the corresponding document sets.

Table 2. Queries and document sets.

Document set id	Query terms (entity concepts)	Document set returned by query
Sd_1	{fox, dog, jumps}	$\{d_1, d_2, d_3\}$
Sd_2	{fox, dog}	$\{d_1, d_2, d_3, d_4\}$
Sd_3	{fox, lazy, dog}	$\{d_1, d_3\}$

Compact: within a document, all groups of concepts (memes) are connected together by common concepts. With reference to Figure 4b, we see that one unique semantic network has been formed by a (weak) link between two memes (concept groups with strong links).

Minimal: each group of concepts (meme) is separated from any other group of concepts. All links (relation concepts) are designated as being strong. With reference to Figure 4b, we see that two distinct memes (concept groups) are identified.

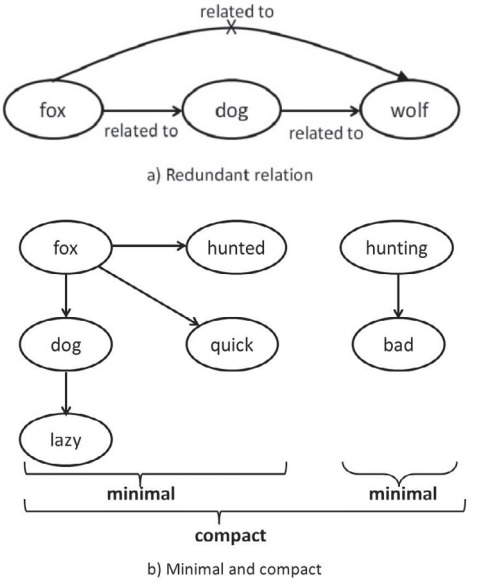

Figure 4. (a) Example of a *redundant* relation in a semantic network; (b) Example of a *compact* and a *minimal* semantic network.

Meme: is a semantic network which is composed of entity concepts with strong links (relation concepts), equivalent to the definition of 'minimal' (above). However, we apply further processing to identify the most relevant memes in a document collection, using the metrics that we will see in the next section.
Incorporation of the Meme Metrics

In this Section we will first describe how we can use the three meme metrics of *longevity*, *fecundity* and *copy-fidelity*, to select the 'top' memes. Then we will give an example of processing using the memes depicted in Figures 2 and 3. We note that we perform the meme metric based selection once the minimal memes have been obtained through the semantic network extraction process described previously.

Meme Metric Based Selection Process
In order to perform the meme metric based selection, we represent the users as a directed graph, through which the memes are considered to 'move'. The implementation details of the graph and associated data structures are out of the scope of the present paper.

The selection process is performed in four steps: ***(i)*** obtain a value for each of the three metrics for each meme; ***(ii)*** obtain the distribution of the values of each metric for all memes; ***(iii)*** establish a cut-off point (threshold) for each metric based on their distributions; ***(iv)*** identify the memes which are above the thresholds for all metrics.

Step 1: obtain a value for each of the three metrics for each meme.
1.1. Longevity *L* for a given meme *m* is designated as m_L. m_L is equal to the number of different arcs that are traversed in a period of time *t*. *Implementation:* this is a simple numerical calculation derived from the initial and maximum timestamps.
1.2. Fecundity *F* for a given meme *m* is designated as m_F. m_F is equal to the number of different vertices that are visited in a given period of time *t*. *Implementation:* this is a simple numerical calculation derived from the directed graph of the users.
1.3. Copy-fidelity *I* for a given meme *m* is designated as m_I. m_I is equal to the degree of 'loss of fidelity' of a meme over a given time period *t* or for a given number of arcs traversed. *Implementation:* a similarity comparison function, with an appropriate distance metric, will be applied to evaluate the fidelity of a given meme (at time *t*) with respect to the original meme (at time *0*).

Note 1, graph structure: the representation of the users and the meme transit between the users will require the implementation of the appropriate data structures and data processing procedures.

Step 2: obtain the ordered distribution of the values of each metric for all memes.
2.1. The distribution of the longevity values m_L for all memes will be a vector d_L. The distribution of the fecundity values m_F for all memes will be a vector d_F. The distribution of the copy-fidelity values m_I for all memes will be a vector d_I.

Step 3: establish a cut-off point (threshold) for each metric based on their distributions:
3.1. The threshold for the longevity distribution d_L will be designated as λ. The threshold for the fecundity distribution d_F will be designated as φ. The threshold for the copy-fidelity distribution d_I will be designated as σ.
Note 2, thresholds: there are different statistical techniques we can use to assign the thresholds λ, φ and σ based on the numerical distribution. For example, we can identify an inflexion point, or we can use the top *x*% percentile, or use a supervised optimization technique. This process could be manual, automatic or semi-automatic.

Step 4: identify the memes which are above the thresholds for all metrics:
4.1. **Meme characteristics.** Consider a meme *m* whose characteristics *mc* are embodied as: concept entities $\{e_1, ..., e_n\}$, concept relations $\{r_1, ..., r_m\}$, longevity m_L, fecundity m_F and copy-fidelity m_I.
4.2. **Meme threshold based selection.** *MT*(*mc*, λ, φ, σ)is a meme threshold selection function, whose inputs for a given meme are the meme's characteristics, mc, as defined in *Step 4.1*, and the three thresholds as obtained from *Steps 1.1* to *3.1*. The output of function *MT* will be a binary value [0,1] for which 1 signifies that meme *m* is within all three thresholds and 0 signifies that it is not. We note that we could relax the meme threshold restrictions, to require only two, or just one threshold to be complied with.

Example of meme metric based selection

In this Section we will give a simple example of the meme threshold based selection, with reference to the memes m_1, m_2 and m_3 shown in Figures 2 and 3. We note that, in this example, time is measured as the number of arcs traversed, and no t the difference between the timestamps.

Applying *Step 1* we obtain:

- Meme longevity: $m_{L1} = 3$, $m_{L2} = 1$, $m_{L3} = 2$
- Meme fecundity: $m_{F1} = 2$, $m_{F2} = 0$, $m_{F3} = 1$
- Copy-fidelity: $m_{I1} = 3$, $m_{I2} = 1$, $m_{I3} = 1$

Applying *Step 2* we obtain the distributions for each metric:

$d_L = \{3, 2, 1\}$; $d_F = \{2, 1, 0\}$; $d_I = \{3, 1, 1\}$

Applying *Step 3 we* establish the threshold for each metric distribution:

$\lambda = 3$; $\varphi = 2$; $\sigma = 3$

Applying *Step 4.1* we assign the meme characteristics for memes m_1, m_2 and m_3, respectively (refer to Figure 3 for the meme definitions and their corresponding concepts):

mc_1({fox, dog}, {jumps-over}, 3, 2, 3); mc_2({fox, dog, brown, quick, lazy}, {is, jumps-over}, 1, 0, 1); mc_3({fox, dog}, {hunts}, 2, 1, 1).

Finally, applying *Step 4.2* identifies the meme(s) which are above the thresholds for all metrics:

$MT(mc_1, \lambda, \varphi, \sigma) = 1$; $MT(mc_2, \lambda, , \varphi, \sigma) = 0$; $MT(mc_3, \lambda, , \varphi, \sigma) = 0$.

Hence, m_1 is the only meme which is above all three thresholds and is therefore selected as the top meme based on the metric thresholds.

CONCLUSIONS

In this paper we have given some formal definitions for memes, in terms of information retrieval and semantic network concepts. We have given some examples which illustrate how these definitions can be used to identify, extract and process memes from an online forum. Then we have used the meme metrics to select the memes in terms of importance and quality, for the given document set. This work lays the ground for future work in which we will process large real online forums containing free text documents (comments), and further develop the formal definitions of memes and their behaviour in different scenarios.

ACKNOWLEDGMENTS

This work was partially funded by Grants TIN2012-38741 (Understanding Social Media: An Integrated Data Mining Approach) of the Ministry of Economy and Competitiveness of Spain, and ARES CONSOLIDER INGENIO 2010 CSD2007-00004.

REFERENCES

[1] R. Dawkins. “The Selfish Gene”. Second edition. Oxford University Press. 1989.

[2] S.J. Blackmore. “The Meme Machine”. Oxford University Press. ISBN: 019286212X. 1999.

[3] S. Ranu, V. Chaoji, R. Rastogi and R. Bhatt. “Recommendations to Boost Content Spread in Social Networks”. World Wide Web 2012. Lyon, France. 2012.

[4] F. Heylighen and K. Chielens. “Cultural Evolution and Memetics”. Article prepared for the Encyclopedia of Complexity and Systems Science, Editors: Robert A. Meyers Ph. D. ISBN: 978-0-387-75888-6 (Print) 978-0-387-30440-3 (Online). 2013. Date of visit: March 15, 2013. URL: http://pespmc1.vub.ac.be/Papers/Memetics-Springer.pdf

[5] G. Bordogna and G. Pasi. “An Approach to Identify Ememes on the Blogosphere”. IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology. Macau, China. 2012.

[6] J. Leskovec, L. Backstrom and J. Kleinberg. “Meme-tracking and the dynamics of the news cycle”. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD ‘09). New York, USA. 2009.

[7] M.P. Simmons, L.A. Adamic and E. Adar. “Memes Online: Extracted, Subtracted, Injected, and Recollected”. Fifth International AAAI Conference on Weblogs and Social Media (ICWSM). Barcelona, Spain. 2011.

[8] D.F. Nettleton. “Data mining of social networks represented as graphs”. Computer

Science Review. Vol. 7, pp. 1-34. February, 2013. DOI: 10.1016/j.cosver.2012.12.001.

[9] A.G. Baydin and R. López de Mántaras. "Evolution of ideas: A novel memetic algorithm based on semantic networks". IEEE Congress on Evolutionary Computation (CEC). Brisbane, Australia. 2012.

[10] S. Szumlanski and F. Gomez. "Automatically acquiring a semantic network of related concepts". 19th ACM Int. Conf. on Information and Knowledge Management. Toronto, Canada. 2010.

[11] G.A. Miller. "WordNet: A Lexical Database for English". Communications of the ACM. Vol. 38, Issue 11, pp. 39-41. 1995.

[12] J. Jiang and D. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy". Int. Conf. on Research in Computational Linguistics. Taipei, Taiwan. 1997.

[13] Z. Chen, A. Gangopadhyay, G. Karabatis, M. McGuire and C. Welty. "Semantic Integration and Knowledge Discovery for Environmental Research". Journal of Database Management. Vol. 18, Issue 1, pp. 43-67. January-March, 2007.

[14] S. Kok and P. Domingos. "Extracting Semantic Networks from Text Via Relational Clustering". European Conf. on Machine Learning and Knowledge Discovery in Databases. Antwerp, Belgium. 2008.

[15] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead and O. Etzioni. "Open information extraction from the web". IJCAI-2007. Hyderabad, India. 2007.

[16] M. Sanderson and B. Croft. "Deriving concept hierarchies from text". 22nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '99). New York, USA. 1999.

[17] J. Oh, T. Kim, S. Park and H. Yu. "PubMed Search and Exploration with Real-Time Semantic Network Construction". 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. Beijing, China. 2012.

[18] R. Ferre-i-Cancho and R.V. Solé. "The Small World of Human Language". Proc. of The Royal Society of London. Series B, Biological Sciences. Vol. 268, pp. 2261-2266. 2001.

[19] J.F. Sowa. "Principles of Semantic Networks: Explorations in the Representation of Knowledge". Morgan Kaufmann. San Mateo. 1991.

[20] R. Mihalcea and D. Radev. "Graphs-Based Natural Language Processing and Information Retrieval". Cambridge University Press. 2012.