International Journal of Computational Intelligence Systems, Vol. 1, No. 1 (January, 2008) 19–23

WORD SIMILARITY FROM DICTIONARIES: INFERRING FUZZY MEASURES FROM FUZZY GRAPHS

Vicenç TORRA[1], Yasuo NARUKAWA[2]

[1] *IIIA, Institut d'Investigació en Intel·ligència Artificial, CSIC, Spanish Council for Scientific Research*
Campus de Bellaterra, 08193 Bellaterra, Catalonia, Spain
E-mail: vtorra@iiia.csic.es

[2] *Toho Gakuen, 3-1-10 Naka, Kunitachi, Tokyo, 186-0004 Japan*
E-mail: narukawa@d4.dion.ne.jp

Received September 10th, 2007
Revised October 5th, 2007

The computation of similarities between words is a basic element of information retrieval systems, when retrieval is not solely based on word matching. In this work we consider a measure between words based on dictionaries. This is achieved assuming that a dictionary is formalized as a fuzzy graph. We show that the approach permits to compute measures not only for pairs of words but for sets of them.

Keywords: Fuzzy measure, Information retrieval, Similarities

1. Introduction

Information retrieval has been a hot research topic in the last years (see *e.g.* [9,3,11]). The Internet and the search engines has increased the need for tools and methods for accessing information in an efficient manner.

When information is textual, users are required to access data from a set of keywords. Such keywords are then matched against inverted indices to retrieve those documents that contain the keywords. Due to the richness of natural language such approach is not always optimal from the point of view of the user. A major point is that language contains synonyms and homophones.

To deal with synonyms, similarity functions can be defined to compare the similarity among two words. Such functions can be either defined on purpose (by the designer of the system) or can be *automatically* extracted from dictionaries (as WordNet) or word corpus. [5] and GAMBAL [11] computed similarities from dictionaries. The systems using Latent Semantics Analysis [4] correspond to the second approach.

In a recent paper [8], we proposed a way to construct fuzzy measures from graphs. Here we review this approach and explore its applicability to define similarity functions between words. A relevant aspect of our work is that it permits us to compute not only the similarity between pairs of words, but also of sets of words or sentences. Such extended similarity is based on a the assumption that exists a basic similarity function already defined on pairs of words.

The structure of this paper is as follows. In Section 2 we review our approximation to construct fuzzy measures from graphs. Then, in Section 3 we describe how this method can be used in information retrieval. The paper finishes with some conclusions and future work.

2. Preliminaries

In this section we review the results that show how a fuzzy measure can be defined from a fuzzy graph. First we review the definition of a fuzzy measure, then, the one of a fuzzy graph and finally the approach to define a fuzzy measure from such fuzzy graph.

Definition 1. A set function $\mu : 2^N \to [0,1]$ is a fuzzy measure if it satisfies the following axioms:

(i) $\mu(\varnothing) = 0$, $\mu(N) = 1$ (boundary conditions);

(ii) $A \subseteq B$ implies $\mu(A) \leqslant \mu(B)$ (monotonicity) for $A,\ B \in 2^N$.

Definition 2. A binary operation $\oplus : [0,\infty) \times [0,\infty) \to [0,\infty)$ is called a pseudo-addition if the following properties are satisfied:

Commutativity: $a \oplus b = b \oplus a$;
Monotonicity: $a \leqslant a'$, $b \leqslant b'$ implies $a \oplus b \leqslant a' \oplus b'$;
Associativity: $(a \oplus b) \oplus c = a \oplus (b \oplus c)$;
Continuity:
$a_n \to a$ and $b_n \to b$ imply $a_n \oplus b_n \to a \oplus b$;
Zero element: $0 \oplus a = a \oplus 0$; for $a,\ b \in [0,\infty)$.

Example. For fixed $p > 0$, let

$$x \oplus y := (x^p + y^p)^{\frac{1}{p}}$$

Then, $\oplus$ is a pseudo-addition.

Now, we define fuzzy graphs. Note that at present several alternative definitions exists for fuzzy graphs, some of them can be found in [1]. See also [10,7] on fuzzy graphs. Roughly speaking, fuzzy graphs have been defined adding fuzziness either on the vertexes or on the edges.

Definition 3. Let N be a finite set and let $\mathscr{R}$ be a fuzzy relation on N (that is, $\mathscr{R} \subset N \times N$, where $\mu_{\mathscr{R}} : N \times N \to [0,1]$ is its membership function), then $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ is a fuzzy graph.

Definition 4. Let $\mathscr{R}$ be a fuzzy relation on N, we say that $T \subset \mathscr{R}$ is a fuzzy tree if there exists no $x_i \in N$ $(2 \leqslant i \leqslant n)$ such that $(x_1,x_2),\ldots,(x_{i-1},x_i) \in \mathscr{R}$ and $x_1 = x_i$. In other words, there are no cycles in T.

We will use $\mathscr{T}_{\mathscr{R}}$ to denote the set of all fuzzy trees of the fuzzy graph $(N, \mathscr{R}, \mu_{\mathscr{R}})$.

Definition 5. Let $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ be a fuzzy graph. Then, we define a set function $m : 2^{\mathscr{R}} \to [0,\infty)$ by:

$$m(A) := \sup_{I \in \mathscr{T}_{\mathscr{R}}} \left\{ \bigoplus_{(x,y) \in I} \mu(x,y) \mid I \subset A \right\} \quad (1)$$

Given the fuzzy graph $\mathscr{G}$ and the set function m, we define another set function $\nu : 2^{\mathscr{R}} \to [0,1]$ as follows:

$$\nu(A) := \frac{m(A)}{m(N)} \quad (2)$$

Note that in this definition, $A \in 2^{\mathscr{R}}$ where $\mathscr{R} \subset N \times N$ and, thus, $A \subset N \times N$. In this way, for example, if $N = \{1,2,3,4,5\}$, A can be e.g. $\{(1,2),(2,3),(1,3)\}$ or $\{(1,2),(2,3)\}$.

The next proposition follows from Definition 5

Proposition 1. [8] *Let $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ be a fuzzy graph and m be the set function defined in Definition 5, then the following conditions hold:*

Boundary condition: *$m(\varnothing) = 0$;*
Monotonicity: *$A \subset B$ implies $m(A) \leqslant m(B)$;*
$\oplus$ ***submodularity:***
$m(A) \oplus m(B) \geqslant m(A \cup B) \oplus m(A \cap B)$.

Proposition 2. [8] *Boundary condition and monotonicity in Proposition 1 as well as the definition of ν in Eq. (2) imply that the set function ν on $2^{\mathscr{R}}$ is a fuzzy measure.*

3. Measuring similarities from a dictionary

In this section we explore the definition of a measure for sets of words. The main idea is to assume that there exists a fuzzy graph that establishes some connections between some pairs of words. This fuzzy graph permits, then, to establish the similarity between pairs of words, and also the similarities between sets of dimensions larger than two. We start considering the similarity between pairs of words.

Definition 6. Let D be a dictionary where N is its set of words, $\mathscr{R}$ be a set of pairs of words that are connected in some sense and let $\mu_{\mathscr{R}}$ be a measure of the strength of the connectivity. Then, D can be expressed as a fuzzy graph $D = (N, \mathscr{R}, \mu_{\mathscr{R}})$. From

now on, to unify the notation with the previous section, and as D is a graph, we will use $\mathscr{G}$ to denote the dictionary.

The definition above can be applied to different types of dictionaries. In particular, if we consider Wordnet [12,2], N is the set of words indexed by Wordnet and $\mathscr{R}$ are the pairs of words that are connected in any of its form of relation. *E.g.* synonyms, hypernyms, hyponyms, *has-part*, etc. Finally, $\mu_{\mathscr{R}}$ is a measure defined on the links in $\mathscr{R}$.

Now, given a dictionary $\mathscr{G}$, and two words w_1 and w_2, we define their similarity as follows:

Definition 7. Let $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ be a fuzzy graph, and let w_1 and w_2 be two elements of N, then, the similarity between w_1 and w_2 denoted $\mathrm{Sim}_{\mathscr{G}}(\{w_1, w_2\})$ is computed as follows:

Let $\mathscr{P}$ be the set of all non-cyclic paths from w_1 to w_2.

Let A be the set of all links that define $\mathscr{P}$. This is, $A = \bigcup_{(x,y)\in\bigcup_{p\in\mathscr{P}} p}(x,y)$.

Then, we define $\mathrm{Sim}_{\mathscr{G}}(\{w_1, w_2\})$ as $\nu(A)$.

Note that $\mathrm{Sim}_{\mathscr{G}}(\{w_1, w_2\})$ measures the strength or similarity between w_1 and w_2 with respect to the whole graph, as $\nu(A)$ contain the strength of all links.

Note that in this definition, when there is no path between w_1 and w_2, the similarity between w_1 and w_2 is zero. This is established in the following proposition:

Proposition 3. *Let $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ be a fuzzy graph, and let w_1 and w_2 be two elements of N, then, when there is no path between w_1 and w_2, $\mathrm{Sim}_{\mathscr{G}}(\{w_1, w_2\}) = 0$.*

Proof. As there is no path, A is empty in Definition 7, and by Definition 5 the measure is zero. □

Under this definition, if there is a single path from w_1 and w_2, then the similarity between w_1 and w_2 is the $\oplus$-combination of the measures of the individual pairs (x,y) that define the path. This is stated below.

Definition 8. Let $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ be a fuzzy graph, and let w_1 and w_2 be two elements of N, we say that there is a connected path from w_1 to w_2 in $\mathscr{R}$ if there exists $(x_1,x_2),(x_2,x_3),\ldots,(x_{i-1},x_i) \in \mathscr{R}$ such that $x_1 = w_1$ and $x_i = w_2$.

Proposition 4. *Let $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ be a fuzzy graph, and let w_1 and w_2 be two elements of N such that there exists a single connected path p from w_1 and w_2. Then, $\mathrm{Sim}_{\mathscr{G}}(\{w_1, w_2\})$ is*

$$\frac{1}{m(N)} \bigoplus_{(x,y)\in p} \mu(x,y)$$

Proof. To prove this proposition, first we consider Definition 7. Therefore,

$$\mathrm{Sim}_{\mathscr{G}}(\{w_1, w_2\}) = \nu(A)$$

where A is the set of all links that define $\mathscr{P}$. In this case, as there is a single path in $\mathscr{P}$, and this path has been denoted by p above, we have that:

$$A = \bigcup_{(x,y)\in \bigcup_{p\in\mathscr{P}} p} (x,y) = \bigcup_{(x,y)\in p} (x,y)$$

Now, we consider $\nu(A)$:

$$\nu(A) = \frac{1}{m(N)} \sup_{I\in\mathscr{T}_{\mathscr{R}}} \left\{ \bigoplus_{(x,y)\in I} \mu(x,y) | I \subset A \right\}$$

Here, as $m(N)$ is constant, will not be considered again.

Now, first recall that the path p from w_1 to w_2 is not cyclic. Thus, $p \in \mathscr{T}_{\mathscr{R}}$. Moreover, all other I such that $I \in \mathscr{T}_{\mathscr{R}}$ will be subpaths of p.

Then, as $\oplus$ is monotonic, $a \oplus 0 = a$ and all $\mu(x,y)$ are positive, we have that the largest value for all I will be obtained considering p. This is:

$$\bigoplus_{(x,y)\in I} \mu(x,y) \leqslant \bigoplus_{(x,y)\in p} \mu(x,y)$$

Therefore, the proposition is proven. □

Note that when several paths can be found in the graph, it is not true that $\mathrm{Sim}_{\mathscr{G}}(\{w_1, w_2\})$ is the $\oplus$ of the measure μ of the links in all the paths.

The definition given above for two words can be easily extended to sets of words. In this case, we first consider paths from any pair of word, and from

these paths we apply the same procedure established in Definition 7.

Definition 9. Let $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ be a fuzzy graph, and let W a subset of N (a set of words), then, the similarity between elements in W denoted $\mathrm{Sim}_{\mathscr{G}}(W)$ is computed as follows:

Let $\mathscr{P}$ be the set of all non-cyclic paths from w_i to w_j where $w_i,\ w_j \in W$.

Let A be the set of all links that define $\mathscr{P}$. This is,

$$A = \bigcup_{(x,y)\in \bigcup_{p\in\mathscr{P}}} p(x,y)$$

Then, we define $\mathrm{Sim}_{\mathscr{G}}(W)$ as $\nu(A)$.

Table 1. Number of adjectives in Wordnet 1.7 with the corresponding number of Synonyms

Num. adjectives	Num. of Synonyms
3487	2
1078	3
404	4
216	5
93	6
49	7
22	8
18	9
15	10
21	11
9	12
5	13
7	14
6	15
2	16
2	17
3	18
1	21
1	22
1	25
1	26
1	27
1	29

Naturally, this definition generalizes Definition 7 when the cardinality of W is two ($W = \{w_1, w_2\}$). Additionally, this definition satisfies the following properties:

Proposition 5. *Let $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ be a fuzzy graph, and let W a subset of N (a set of words), then:*

1. *If for all $w_i,\ w_j$ in W, there is no path $\mathscr{G}$ connecting them, then $Sim_{\mathscr{G}}(W) = 0$.*
2. *If $W = N$ (all the words are considered), $Sim_{\mathscr{G}}(W) = 1$.*

Proof. The proof of the first boundary condition is similar to the one in Proposition 3: As no path exists, the measure becomes zero.

The proof of the second boundary condition is based on the fact that when $W = N$, all paths are visited and, thus, all links are considered. As each link is only visited once in $A = \bigcup_{(x,y)\in\bigcup_{p\in\mathscr{P}}} p(x,y)$, they are the same links visited by $m(N)$. Therefore, the outcome is one. □

Proposition 6. *Let $\mathscr{G} = (N, \mathscr{R}, \mu_{\mathscr{R}})$ be a fuzzy graph, and let W_1 and W_2 subsets of N (sets of words) such that $W_1 \subseteq W_2$, then*

$$Sim_{\mathscr{G}}(W_1) \leqslant Sim_{\mathscr{G}}(W_2)$$

This is, the similarity is monotonic with respect to the set of words.

3.1. Computational issues

According to Definitions 7 and 9, the approach presented here requires the computation of $m(N)$. Taking into account the proof of Proposition 5, we have that all links are only considered once in $m(N)$, therefore computation of $m(N)$ only requires the determination of the number of links in the dictionary and its weight.

For example, in the case of Wordnet [12], this can be done through inspection of the files. In the case of the adjectives, the file "index.adj" can be used to count the number of meanings (synsets) of one adjective. Each meaning can then be further linked with several words through the "data.adj" file. In the particular case of Wordnet 1.7 there are 18523 synsets and 22495 synonyms established between synsets in this latter file.

The relationships between adjectives and synsets in such version of Wordnet are as follows. There are 21359 adjectives, of which 5443 include at least two synsets in their meaning. The exact number of adjectives/synsets is included in Table 1. From this table, we can determine that the number of links between adjectives and synsets is:

$$15154 + (21359 - 5443) = 31070$$

So, all together, the graph for the adjectives, representing only synonymy contains $22495 + 31070 = 53565$ links.

Assessing a measure m equal to α for links between synsets and β for links between adjectives and synsets, $m(N)$ would be equivalent, if only adjectives are considered to:

$$m(N) = 22495\alpha + 31070\beta$$

Although this graph is huge, the computation of the similarity measure for a pair of words (or of a set of words) is computationally similar to the approaches in [5,11] for most situations. Note that in these latter cases, all paths between the words considered should be detected. In the approach proposed here, we also need to determine these paths. Differences correspond to the way the computation of m is done. The approach presented here has the advantage of being usable when the similarity between sets of more than two words are considered.

4. Conclusions and future work

In this paper we have studied the definition of similarity measures between words based on dictionaries. We have shown that when a dictionary is represented as a fuzzy graph, we can build a fuzzy measure that can be used to compute a degree of similarity between sets of words. We have considered the application of our approach to the case of Wordnet.

As future tasks to be accomplished, we consider the implementation of the approach in our system GAMBAL, and to consider whether simpler equivalent expressions can be found for computing the similarity for sets of words.

Acknowledgements

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) and grant "Salvador de Madariaga" PR2007-0122 is acknowledged.

References

1. M. Blue, B. Bush, J. Puckett, "Unified approach to fuzzy graph problems", *Fuzzy Sets and Systems*, **125**, 355–368 (2002).
2. C. Fellbaum, WordNet: *An Electronic Lexical Database*, The MIT Press, (1998).
3. E. Herrera-Viedma, "Fuzzy Qualitative Models to Evaluate the Quality on the Web", (MDAI 2004), *Lecture Notes in Artificial Intelligence*, **3131**, 15–26 (2004).
4. LSA (2007) http://lsa.colorado.edu/
5. R. Mandala, T. Tokunaga, H. Tanaka, "Query expansion using heterogeneous thesaury", *Information Processing and Management*, **36**, 361–378 (2000).
6. S. Miyamoto, K. Mizutani, "Fuzzy Multiset Model and Methods of Nonlinear Document Clustering for Information Retrieval", (MDAI 2004), *Lecture Notes in Artificial Intelligence*, **3131**, 273–283 (2004).
7. J.N. Mordeson, P.S. Nair, *Fuzzy Graphs and Fuzzy Hypergraphs*, Physica-Verlag, Heidelberg, (2000).
8. Y. Narukawa, V. Torra, "Choquet integral on discrete spaces", in: B. Reusch (Ed.), *Computational Intelligence, Theory and Applications*, (ISBN 3-540-22807-1, Heidelberg: Physica-Verlag Springer, series on 'Advances in Soft Computing'), 573–581 (2005).
9. G. Pasi, "Modeling users' preferences in systems for information access", *Int. J. of Intelligent Systems*, **18** (7), 793–808 (2003).
10. A. Rosenfeld, "Fuzzy graphs", in: L.A. Zadeh, K.S. Fu, K. Tanaka, M. Shimura (Eds.), *Fuzzy Sets and their Applications to Cognitive and Decision Processes*, Academic Press, New York, 77–95 (1975).
11. V. Torra, S. Miyamoto, S. Lanau, "Exploration of textual databases using a fuzzy hierarchical clustering algorithm in the GAMBAL system", *Information Processing and Management, Information Processing and Management*, **41** (3), 587–598 (2005).
12. Wordnet (2007), http://wordnet.princeton.edu/