A Novel View on Information Content of Concepts in Extremely Large Ontologies

Carl Van Buggenhout[a], Werner Ceusters[a]

[a]*Language and Computing nv., Zonnegem, Belgium*

Abstract

Semantic distance and semantic similarity are two important information retrieval measures used in word sense disambiguation as well as for the assessment of how relevant concepts are with respect to the documents in which they are found. A variety of calculation methods have been proposed in the literature, whereby methods taking into account the information content of an individual concept outperform those that don't. In this paper, we present a novel recursive approach to calculate a concept's information content based on the information content of the concepts to which it relates. The method is applicable to extremely large ontologies containing several million concepts and relationships amongst them. It is shown that a concept's information content as calculated by this method provides additional information with respect to an ontology that cannot be approximated by hierarchical edge-counting or human insight. In addition, it is suggested that the method can be used for quality control within large ontologies.

Keywords:

Semantic Distance, Information Content

1 Introduction

Semantic distance and semantic similarity are important measures in concept-based information retrieval. They have been used with varying degrees of success in applications such as medical coding [1], semantic indexing [2], word sense disambiguation [3, 4] and image caption retrieval [5] to mention only a few. Whereas semantic distance measures how closely two concepts are topologically related in a semantic network, semantic similarity captures to what extent two concepts might represent the same thing. Obviously, the two notions are closely related, although not the same. A concept such as "fractured arm" should have a very short semantic distance towards "arm fracture", whereas the semantic similarity should be small: a "fracture" cannot stand for an "arm" or the other way round. But, to complicate matters, any descent system should be able to compute the semantic distance of post-coordinated concepts such as "patient"-WITH-"arm fracture" and "patient"-WITH-"fractured arm" as being minimal, and the semantic similarity as being maximal.

Various approaches to calculate both values have been proposed. They tend to fall in two categories. *Edge-based methods* exploit mainly the idea of path-length in a network with or without additional weights according to the type of link traversed, whereas *node-based methods* also take into account the probability to find each concept in a large corpus [6]. The idea behind them is that the "information content" of concepts occurring often in a corpus is lower than of concepts that occur rarely, and that these information-low concepts tend to appear higher in an ontology. Recently, a similar idea is introduced in methods that are intrinsically edge-based. It tends to capture the feeling that the semantic difference between upper level concepts in an ontology is bigger than between lower level concepts [7]. The implementation in [7] is entirely based on the hierarchical ISA-relationship. In this paper, we expand this idea by taking also into account the associative relationships amongst concepts.

2 Material and methods

LinKBase® is a large scale medical ontology developed and maintained by the modeling team of Language and Computing nv. LinKBase® contains currently over one million language-independent medical and general-purpose concepts, linked to natural language terms in several languages, including English [8, 9]. These concepts are linked together into a semantic network like structure using approximately 450 different link types for expressing formal relationships. These relationships are based on logics dealing with issues such as mereology and topology [10, 11], time and causality [12] and models for semantics driven natural language understanding [13, 14]. Link types form a multi-parented hierarchy on their own. It is very important to note that in LinkBase® the formal subsumption relationship covers only about 15% of the total number of relationships amongst concepts. As such, LinkBase® is a much richer structure than terminological systems in which term-relationships are expressed as strictly "narrower" or "broader". Important to note also is that LinkBase® is a "living" ontology, in which data are changed on a daily basis, and at a rate of 2000 to 4000 modifications a day. Moreover, it is not required for concepts added to be perfectly modeled from the very beginning [15].

We defined the initial information content (IC_0) of a concept in LinkBase® as:

$$\forall k : IC_0(C_k) = 1 + \sum_i (LW(L_i)*IC_0(C_i)) \qquad (1)$$

where C_k is the source concept, C_i the target concept, $LW(L_i)$ the link weight of the link between C_k and C_i, and the sum is going over all the outgoing concepts of C_k.

We defined the initial link weight (LW_0) of a link type as:

$$\forall k : LW_0(L_k) = 1 + \sum_i (LW_0(ParentLink_i(L_k))) \qquad (2)$$

where the sum goes over all the parents of link type L_k.

From these formulas it follows that the IC of individual concepts can only be found by setting up and subsequently solving a (huge !) system of equations, i.e. one per concept in the ontology. Once all values computed, they were normalised using a straight line in the range of 0 to 1. We used the following function for the IC's:

$$IC := (C, a) \rightarrow \frac{IC_0(C)}{a} \text{ where } a = \max(IC_0(C_k)), \forall k \qquad (3)$$

and we used this function to normalize the LW's:

$$LW := (L, a) \rightarrow \frac{LW_0(L)}{a} \text{ where } a = \max(LW_0(L_k)), \forall k \qquad (4)$$

An algorithm was designed to compute the IC's in real time, taking advantage of the network-structure of the ontology.

A first analysis of the results was carried out by comparing the information content ranking of the algorithm with those of human judges. All concepts containing the substring “tachycard” as part of their knowledge name were extracted, ranked alphabetically, and given to the ontology modelers (medical doctors with at least one year experience in ontology building). Their task was to rank the concepts according to their subjective information content. The results of the evaluators were then compared with those of the algorithm. The null-hypothesis for this analysis is that there would be no difference between human judges as a group, and the automatic ranking based on IC.

A second analysis involved a comparison of the depth of a concept in the hierarchy with respect to its computed IC value. For the IC to fulfill its purpose, there should be no complete correlation.

3 Results

The program ran on an Intel Pentium 2.4 GHz processor and used approximately 600 MB of RAM memory. It took about 15 minutes to calculate the 450 link-type weights and the 1 million IC’s, and also some 10 minutes to store all the concept values in a file, sorted on IC.

3.1 Evaluation of the ranking experiment

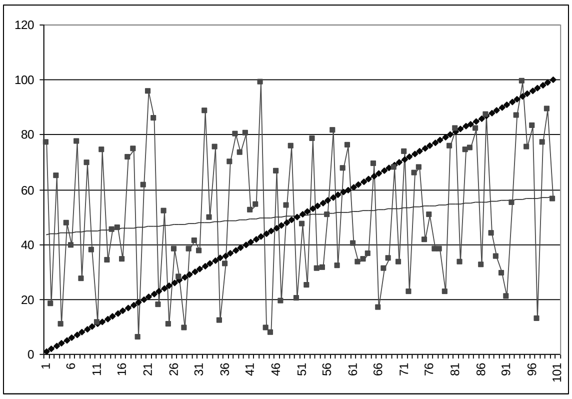

Figure 1: IC-ranking by algorithm and human evaluators

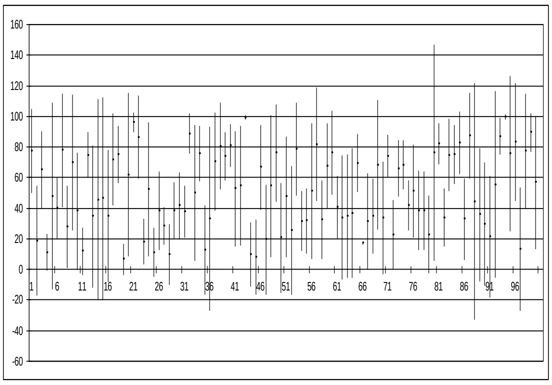

Figure 2: inter-rater agreement amongst human evaluators

Figure 1 captures the results of the ranking performed by the algorithm (diagonal line with diamonds) and the human evaluators (irregular line with squares, based on the overall mean of the different human ratings). The third, unmarked line, is the result of a trend analysis performed on the human ratings. From this it follows that the upward trend of the ranking produced by the algorithm is followed by the human raters, but at a considerable lower slope.

Figure 2 shows the mean rankings of the human raters within a 2 standard deviation confidence interval. From this it follows that the human raters made quite different assessments of the information content.

3.2 Correlation between the hierarchical depth of a concept in the ontology and its IC

Figure 3 shows a scatter plot of the normalized IC’s of the “tachycardia”-related concepts, versus the depth according to the ISA-hierarchy. Because the normalised IC’s are very small, they were re-scaled using formula 5, for pure visualisation purposes.

$$\text{rescaled IC} = -100/\log(\text{IC}) \qquad (5)$$

As can be seen, the hierarchical depth of a concept in the ontology is not the most important factor that contributes to the information content.

Figure 4 gives an impression concerning the distribution of the normalized IC's on a logarithmic scale. Only the first 100 concepts with the highest IC are shown.

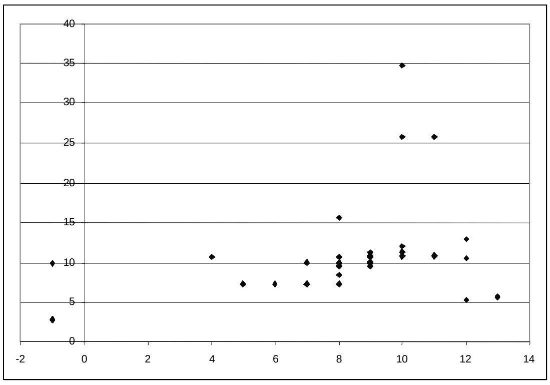

Figure 3: IC/depth plot

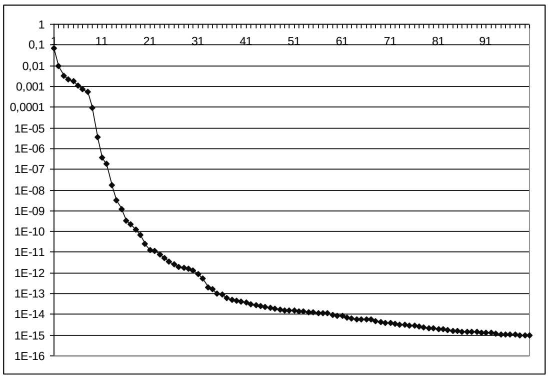

Figure 4: IC distribution

4 Discussion

4.1 Algorithm design

To assess how much information a concept contains about other concepts, we had to set up a system of equations. This system was set up by assuming that the information content of a concept relates to the sum of the IC's of the concepts used to describe it, weighted by a link type specific value (formula (1)). These link type specific values were calculated in exactly the same way, but without a weighting factor (formula (2)).

A naïve approach to set up the system of equations would be to generate the equation for each concept individually and then to use a matrix formalism to calculate the IC for each concept. However, given the huge size of the ontology, this would not be sensible.

We developed a novel algorithm exploiting the fact that generation and partially solving the equations could be guided by the structure of the ontology.

There are two steps in the algorithm. In the first step LW's are calculated that are used in the second step to compute the IC's of the concepts.

We start the work with a list of all the link types. For every link type L1 that we didn't visit yet, we calculate its LW by following all the parent link types of this link type, and using their LW's to compute the value of L1 according to formula (2). If the LW of one of the outbound link types (say L2) isn't computed up till now, we calculate the value for link type L2 by recursively applying the same procedure that we used for calculating the LW of L1, that is, we follow all the parent link types of link type L2, and use their LW's to compute the value of L2, until we reach a link type Li that has no parents. The value for this link type Li is 1. After we calculated all the LW's, we normalize them using formula (4).

We then work with a list of all the concepts. For every concept C1 that we didn't visit yet, we calculate its IC by following all the outbound links of this concept, and using the concept values of these outbound concepts to compute the value of C1 according to formula (1). If the concept value of one of the outbound concepts (say C2) isn't computed up till now, we calculate the value for concept C2 by recursively applying the same procedure as for calculating the concept values of C1, that is, we follow all the outbound links of concept C2, and use their IC's to compute the value of C2. This process comes to an end if we reach a concept Ci that has no parents. Its value equals 1. After we calculated all the IC's, we normalize them using formula (3).

There is one condition that must be taken into account: the algorithm we invented iterates until it reaches a value that has already been calculated, so this value has not to be computed for a second time. However when there are cycles – i.e. paths from a concept towards the same concept – in the ontology, we must stop the iteration process for instance after n iterations. We are sure that the value of such a concept converges to its 'real' value if the system of equations is not false. In that case the more iterations the program does, the more precise these converging values get.

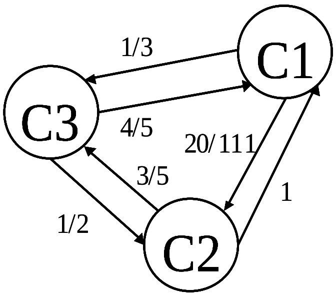

Figure 5: unsolvable configuration

But what if the system is false? With values the 6 links as in figure 5, the system of equations is not solvable. This is because the determinant of the coefficient-matrix of the system equals zero. Therefore we can't use algorithms that make use of the inverse of that matrix because the inverse doesn't exist. Such algorithms are for instance Gauss-Seidel (GS) or SOR (successive over relaxation) [16]. Both algorithms assume that the coefficient-matrix is diagonally. The more diagonally the matrix, the faster the solution converges. But they fail on situations as in Figure 5. Our algorithm handles this situation by calculating the best approximation.

4.2 Interpretation of the results

It was no surprise to us that the modelers' rankings differed considerably amongst each other, as well as with respect to the algorithm's ranking. Differences amongst modelers could be explained most often by inaccurate estimations of the IC of additional criteria associated to the concept that appeared to be more central. The IC of the concept "chronic tachycardia" was by all modelers correctly judged lower than the IC of "fetal tachycardia", which on its turn was judged lower than for "fetal tachycardia affecting management of mother". But typically, the IC-differences for "chronic", "fetal", and "affecting management of mother" were seriously underestimated. When informed about these differences, some modelers accepted this view without critique, while others judged the differences as real but irrelevant, a situation similar as in [17] where only 2 out of 19 hierarchic relationships generated by a description logic classifier (hence mathematically correct) were judged "accurate" enough by human reviewers to be taken into account.

It was also no surprise that the IC of a concept is relatively (though not complete of course) independent from it's place in the hierarchy. For a given hierarchical depth, the range of IC's is typically large. However, when standard statistical techniques for outlier detection were used, the majority of outliers turned out to be the result of inappropriate modeling. As such, this method might be useful for quality control.

5 Conclusion

We have been able to design a novel algorithm to calculate the information content of concepts in extremely large ontologies. The method adds another dimension to the notions of semantic distance and semantic similarity as the calculated IC's are relatively independent from the hierarchical depth within an ontology. Because information content has been shown to be an important parameter for accurate information retrieval [5, 7], our method might give an important contribution in that field. In addition, we have indications that the method can also be used for quality control.

6 References

[1] Bousquet C, Jaulent M, Chatellier G, Degoulet P, *Using Semantic Distance for the Efficient Coding of Medical Concepts*. Proceedings of AMIA 2000, http://medicine.ucsd.edu/F2000/D200437.htm

[2] Jackson B, Ceusters W. *A novel approach to semantic indexing combining ontology-based semantic weights and in-document concept co-occurrences.* In Baud R, Ruch P. (eds) EFMI Workshop on Natural Language Processing in Biomedical Applications, 8-9 March, 2002, Cyprus, 75-80.

[3] Sussna M., *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network*, in Proceedings of the Second International Conference on Information and knowledge Management. Arlington, Virginia USA. 1993, 67-74.

[4] Agirre E and Rigau G. *Word Sense Disambiguation Using Conceptual Density.* In: Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, 1996, 16-22.

[5] Smeaton AF and Quigley I. *Experiments on Using Semantic Distances Between Words in Image Caption retrieval*, , in: Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR96) Zurich, Switzerland, 1996, 174-180,

[6] Polčicová, G., and Návrat, P.: *Semantic Similarity in Content-based Filtering*: In proc. of ADBI2002 Advances in Databases and Information Systems, Manolopoulos, Y. and Návrat, P. (Eds.), Springer LNCS 2435, 2002, 80-85.

[7] Zhong J, Zhu H, Li J, Yu Y. *Conceptual Graph Matching for Semantic Search*. In Priss U, Corbett D, Angelova G (eds.) Conceptual Structures: Integration and Interfaces (ICCS2002), 2002, 92-106.

[8] Ceusters W, Martens P, Dhaen C, Terzic B, *LinkFactory: an Advanced Formal Ontology Management System*. Interactive Tools for Knowledge Capture Workshop, KCAP-2001, October 20, 2001, Victoria B.C., Canada (http://sern.ucalgary.ca/ksi/K-CAP/K-CAP2001/).

[9] Montyne F, *The importance of formal ontologies: a case study in occupational health.* OES-SEO2001 International Workshop on Open Enterprise Solutions: Systems, Experiences, and Organizations, Rome, 14-15 September 2001 (http://cersi.luiss.it/oesseo2001/papers/28.pdf).

[10] Smith B, *Mereotopology: a theory of parts and boundaries*, Data and Knowledge Engineering 20 (1996), 287-301.

[11] Smith B, Varzi AC, *Fiat and Bona Fide Boundaries*, in Proc. COSIT-97, Springer-Verlag 1997, 103-119.

[12] Buekens F, Ceusters W, De Moor G, *The Explanatory Role of Events in Causal and Temporal Reasoning in Medicine*, Met Inform Med 1993, 32: 274 - 278.

[13] Ceusters W, Buekens F, De Moor G, Waagmeester A, *The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition.* Met Inform Med 1998; 37(4/5):327-33.

[14] Bateman JA. *Ontology construction and natural language*. In Proc. International Workshop on Formal Ontology. Padua, Italy, 1993, 83-93.

[15] Flett A, Casella dos Santos M, Ceusters W. *Some Ontology Engineering Processes and their Supporting Technologies*, in: Gomez-Perez A, Benjamins VR (eds.) Ontologies and the Semantic Web, EKAW2002, Springer 2002, 154-165.

[16] Hackbusch W. *Iterative Solution of Large Sparse Systems of Equations*. Applied Mathematical Sciences, Vol. 95. Springer-Verlag, New York, U.S.A., 1993.

[17] Hardiker NR, Rector AL, *Structural validation of nursing terminologies*. J Am Med Inform Assoc 8(3), 2001, 212-21.