A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness (2009)

George Tsatsaronis and Vicky Panagiotopoulou

Department of Informatics Athens University of Economics and Business, 76, Patision Str., Athens, Greece

Report by:

Tarun Raheja

2015A7PS0106H

PROBLEM

This paper attempts to develop a new "Generalized Vector Space Model" that broadens the existing Vector Space Model by incorporating additional information, other than just terms, in the document representation. More precisely, it aims to include semantic information about the terms from concept thesauri like WordNet so that word meanings can be inferred when appropriate. The central (and most daunting) task is to accurately extend the existing VSM to have semantic data while maintaining absolute mathematical rigor and accurate outputs. For this purpose, the paper proposes a new heuristic for evaluating semantic relevance. In-depth testing is later used to establish a noticeable improvement in the IR system performance using the aforementioned model.

The literature on usage of semantic data to improve text retrieval performance before this paper was published is abundant, but their results are far from conclusive. While some publications (Mavroeidis, Krovetz and Croft, Vorhees) stipulate that GVSM leads to better overall performance, higher recall and improves results, others (Sanderon) starkly oppose these results and maintain that

there is no guarantee that even highly accurate semantic datasets improve performance. So the issue remains an open problem because of the inability of researchers in the field to agree on a final conclusion. This paper takes the former stance, i.e. sense based IR actually helps in increasing the output quality and recall of the system using their proposed heuristic for semantic relatedness. They propose to incorporate term-term relatedness into the VSM by using the open thesaurus WordNet, and they have a three-pronged stool of conditionals upon which rests the quality of their system. They are:

a) A novel idea for measuring semantic closeness that exploits attributes present in WordNet such as sense depth.
b) A full-fledged mathematical model that wholly incorporates the above heuristic and appends it to the existing VSM model to form the new GVSM model.
c) Complete use of all information offered by the thesaurus to utilize word sense to its full potential, using data such as semantic closeness across parts-of-speech.

The paper then goes on to present basic overviews of VSM and GVSM, and offers contrasts and improvements. It starts off by describing the standard VSM, and how it calculates similarity based on cosine distance between two vectors. It then highlights its key pitfalls, like the assumption that all term vectors are orthogonal, which signifies that there is no significant synonymity within the vocabulary of a language which is obviously a very idealist and unrealistic assumption which leads to quite a few problems.

SOLUTION APPROACH

It then gives a few important characteristics of the GVSM, and how it overcomes these pitfalls. For starters, the GVSM assumes a completely different vector space, which has 2^n vectors instead of the VSM vector space, and hence helps do away with the pairwise orthogonality assumption while still keeping the vectors linearly independent. Now each term and document vector can be represented in terms of the basis vectors of the new space, and the similarity measure then slightly changes and can be evaluated by the equation below:

$$cos(\vec{d_k}, \vec{q}) = \frac{\sum_{j=1}^{\acute{n}} \sum_{i=1}^{\acute{n}} \acute{a_{ki}} \acute{q_j} \vec{t_i} \vec{t_j}}{\sqrt{\sum_{i=1}^{\acute{n}} \acute{a_{ki}}^2 \sum_{j=1}^{\acute{n}} \acute{q_j}^2}}$$

It is important to note that if the term vectors are assumed to be orthogonal, this formula simplifies the vector product to 1 and yields the exact same formula as the VSM.

After this, the paper tries to break down the GVSM into its essential components and briefly elaborate all of them. Semantic closeness evaluation has two strong cornerstones :

1) Computation of meaning similarity.
2) Computation of frequency of the terms occurring together statistically from large corpora.

The paper approaches the problem from the first pathway and shows the approaches taken by earlier papers along with a brief outline of their results. It then explains how the method they used is different from those used earlier, in the sense that it employs semantic closeness as an appendage to the terms, instead of directly replacing the terms with their meanings and synonyms.

Then the paper begins to delve into the computational aspects and the technicalities.

ARCHITECTURE, SOLUTION AND TECHNICAL MINUITAE

The paper points out that many words can have the same meaning (synonymy) and the same word can have several meanings (polysemy). It stipulates that both of these are included in the term product in the equation shown above, and goes on to systematically establish the same.

The technical details are clearly explained in the paper in five distinct parts as follows:

1) **Semantic Relatedness**

Semantic closeness of two words is associated to two newly defined attributes which are taken from WordNet, namely 'compactness' and 'semantic path elaboration'. These terms are then explained in this section. Compactness is defined as the product of edge weights along the path from senses of one term to the other in WordNet. Edge weights are assigned based on how closely two terms are with respect to meaning. It is clarified that compactness can have multiple values based on what path is taken from one sense to the other. Then the other factor of depth is expounded on, which essentially is a measure of how general a path is. A path with nodes that are closer to the surface is considered more general than a path which has deeper nodes.

Then the paper talks about SPE, Semantic Path Elaboration, which is another measure of relatedness. The SPE is heuristically defined as the harmonic mean of the two paths' depths after normalizing with the global thesaurus maximal depth. These two terms are then combined to explain Semantic Relatedness.

SR is defined as the maximum {compactness*SPE} path between two respective senses of two words, which are computed between all pairs of senses for both the words.

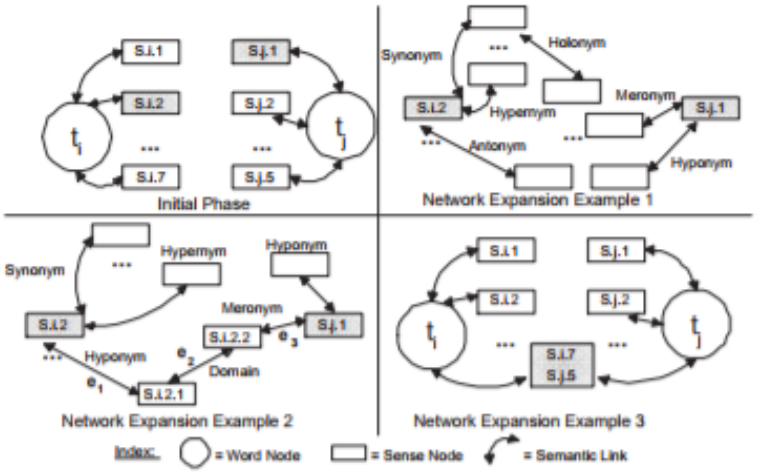

Figure 1: Computation of semantic relatedness.

2) **Semantic Networks from a thesaurus**

The paper exploits a previously used method of network construction from a paper by Tsatsaronis, which utilizes all possible information that WordNet and other thesauri offer. It then takes the reader through a worked example of a semantic relatedness computation.

3) Maximum SR

This section forms the heart of the paper and outlines the algorithms and central theorem that is at the centre of the GVSM.

It is natural that the path between to words should be weighed with the maxima of the computed SR. This computation is done via a proposed Dijkstra-like algorithm:

Algorithm 1 MaxSR(G,u,v,w)

Require: A directed weighted graph G, two nodes u, v and a weighting scheme $w : E \rightarrow (0..1)$.

Ensure: The path from u to v with the maximum product of the edges weights.

Initialize-Single-Source(G,u)

1: **for all** vertices $v \in V[G]$ **do**
2: $d[v] = -\infty$
3: $\pi[v] = NULL$
4: **end for**
5: $d[u] = 1$

Relax(u, v, w)

6: **if** $d[v] < d[u] \cdot w(u, v)$ **then**
7: $d[v] = d[u] \cdot w(u, v)$
8: $\pi[v] = u$
9: **end if**

Maximum-Relatedness(G,u,v,w)

10: Initialize-Single-Source(G,u)
11: $S = \emptyset$
12: $Q = V[G]$
13: **while** $v \in Q$ **do**
14: s = Extract from Q the vertex with max d
15: $S = S \cup s$
16: **for all** vertices $k \in$ Adjacency List of s **do**
17: Relax(s,k,w)
18: **end for**
19: **end while**
20: return the path following all the ancestors π of v back to u

The paper then proves mathematically the correctness of this algorithm. It does so by showing that SR would indeed be the maximal product of all paths leading from one word to the other.

4) **Meaning Disambiguation**

The paper weakly stresses on the fact that a good Word Sense Disambiguation algorithm can accomplish the same task, and then shows how their algorithm does the same task, but quantitatively.

5) **The GVSM Model**

This section puts all the previous definitions together into one neat equation. The crux of it is that since the term vectors in the original GVSM equation are unknown, their vector product can be computed using the above SR algorithm. The similarity scores are then added to the TF IDF scores, and this gives a model in which the terms need not e exact matches, but can be synonymous.

$$cos(\vec{d_k}, \vec{q}) = \frac{\sum_{i=1}^{n} \sum_{j=i}^{n} d_k(t_i, t_j) \cdot q(t_i, t_j)}{\sqrt{\sum_{i=1}^{n} \sum_{j=i}^{n} d_k(t_i, t_j)^2} \cdot \sqrt{\sum_{i=1}^{n} \sum_{j=i}^{n} q(t_i, t_j)^2}} \quad (4)$$

RESULTS

The model is tested completely from the inside out. Firstly the semantic relatedness measure is evaluated using three benchmark data sets, and then the GVSM model is made to retrieve data from three benchmark datasets to accurately evaluate performance.

The authors evaluated their closeness measure against ten standard 'already-known' closeness measures, and computed the Spearman correlation of each of them, and their own with human-judged answers. The method surpasses all other tested methods in all but one dataset, where it is beaten by one of the older methods. This shows that their model is the closest to human word semantic understanding among all the others.

Then the GVSM model was tested on three vastly varying TREC collections. For contrast, the VSM model was used on the same datasets. The resulting precision and recall from the two was compared, and it was clearly seen that the GVSM

model yielded boosts of 0.4% to as much as 1.92% on aforementioned datasets. This clearly shows that even very primitive SR analysis provides significantly better results, and this seems to be quite a promising endeavor for further research.

IMPROVEMENTS

The authors have not used any sort of Natural Language Processing Word Sense Disambiguation algorithms, or equivalent methods to map the terms themselves to their correct sense in the first place. Implementation of this will clearly result in a very strong boost to the model's retrieval power, since the foundational input for semantic analysis will be much more accurate.

Another probable idea is to try attacking the problem from a more phrase-ish point of view. Phrase lookup mechanisms would drastically reduce ambiguity, and hence would make it easier for the retrieval system to associate sense with words since frequencies are known.

Other ideas inspired from other papers (Voorhees) go along the lines of creating another dual space for semantics and using this along with the term space to dual-represent the document. This would be immensely helpful in terms of solving other problems in the field too.

CONCLUSION

The paper successfully established a new heuristic for semantic relatedness computing that performs much better than all the hitherto known metrics. They then embedded this into the existing VSM model to yield the new GVSM model. The SR measure proposed uses WordNet essentially as a graph, computes edge weights based on node type and depth, and then finds the maximum relatedness

between two nodes connected via one or more paths. Then the paper looks into conditions needed for semantic assisted retrieval to be more accurate than VSM and proposes further researchable avenues.