Measuring Conceptual Similarity by Spreading Activation over Wikipedia's Hyperlink Structure

Stephan Gouws, G-J van Rooyen, and Herman A. Engelbrecht
Stellenbosch University
{stephan,gvrooyen,hebrecht}@ml.sun.ac.za

Abstract

Keyword-matching systems based on simple models of semantic relatedness are inadequate at modelling the ambiguities in natural language text, and cannot reliably address the increasingly complex information needs of users. In this paper we propose novel methods for computing semantic relatedness by spreading activation energy over the hyperlink structure of Wikipedia. We demonstrate that our techniques can approach state-of-the-art performance, while requiring only a fraction of the background data.

1 Introduction

The volume of information available to users on the World Wide Web is growing at an exponential rate (Lyman and Varian, 2003). Current keyword-matching information retrieval (IR) systems suffer from several limitations, most notably an inability to accurately model the ambiguities in natural language, such as synonymy (different words having the same meaning) and polysemy (one word having multiple different meanings), which is largely governed by the context in which a word appears (Metzler and Croft, 2006).

In recent years, much research attention has therefore been given to *semantic* techniques of information retrieval. Such systems allow for sophisticated semantic search, however, require the use of a more difficult-to-understand query-syntax (Tran et al., 2008). Furthermore, these methods require specially encoded (and thus costly) *ontologies* to describe the particular domain knowledge in which the system operates, and the specific interrelations of concepts within that domain.

In this paper, we focus on the problem of computationally estimating similarity or relatedness between two natural-language documents. A novel technique is proposed for computing semantic similarity by spreading activation over the hyperlink structure of Wikipedia, the largest free online encyclopaedia. New measures for computing similarity between individual concepts (**inter-concept similarity**, such as "France" and "Great Britain"), as well as between documents (**inter-document similarity**) are proposed and tested. It will be demonstrated that the proposed techniques can achieve comparable inter-concept and inter-document similarity accuracy on similar datasets as compared to the current state of the art Wikipedia Link-based Measure (WLM) (Witten and Milne, 2008) and Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) methods respectively. Our methods outperform WLM in computing inter-concept similarity, and match ESA for inter-document similarity. Furthermore, we use the same background data as for WLM, which is less than 10% of the data required for ESA.

In the following sections we introduce work related to our work and an overview of our approach and the problems that have to be solved. We then discuss our method in detail and present several experiments to test and compare it against other state-of-the-art methods.

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 46–54,

2 Related Work and Overview

Although Spreading Activation (SA) is foremost a cognitive theory modelling semantic memory (Collins and Loftus, 1975), it has been applied computationally to IR with various levels of success (Preece, 1982), with the biggest hurdle in this regard the cost of creating an associative network or knowledge base with adequate conceptual coverage (Crestani, 1997). Recent knowledge-based methods for computing semantic similarity between texts based on Wikipedia, such as Wikipedia Link-based Measure (WLM) (Witten and Milne, 2008) and Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007), have been found to outperform earlier WordNet-based methods (Budanitsky and Hirst, 2001), arguably due to Wikipedia's larger conceptual coverage.

WLM treats the anchor text in Wikipedia articles as links to other articles (all links are treated equally), and compare concepts based on how much overlap exists in the out-links of the articles representing them. ESA discards the link structure and uses only the text in articles to derive an explicit concept space in which each dimension represents one article/concept. Text is categorised as vectors in this concept space and similarity is computed as the cosine similarity of their ESA vectors. The most similar work to ours is Yeh (2009) in which the authors derive a graph structure from the inter-article links in Wikipedia pages, and then perform random walks over the graph to compute relatedness.

In Wikipedia, users create links between articles which are seen to be related to some degree. Since links relate one article to its neighbours, and by extension to their neighbours, we extract and process this hyperlink structure (using SA) as an **Associative Network (AN)** (Berger et al., 2004) of concepts and links relating them to one another. The SA algorithm can briefly be described as an iterative process of propagating real-valued energy from one or more source nodes, via weighted links over an associative network (each such a propagation is called a *pulse*). The algorithm consists of two steps: First, one or more pulses are triggered, and second, termination checks determine whether the process should continue or halt. This process of activating more and more nodes in the network and checking for termination conditions are repeated pulse after pulse, until all termination conditions are met, which results in a final activation state for the network. These final node activations are then translated into a score of relatedness between the initial nodes.

Our work presents a computational implementation of SA over the Wikipedia graph. We therefore overcome the cost of producing a knowledge base of adequate coverage by utilising the collaboratively-created knowledge source Wikipedia. However, additional strategies are required for translating the hyperlink structure of Wikipedia into a suitable associative network format, and for this new techniques are proposed and tested.

3 Extracting the Hyperlink Graph Structure

One article in Wikipedia covers one specific topic (concept) in detail. Hyperlinks link a page A to a page B, and are thus *directed*. We can model Wikipedia's hyperlink structure using standard graph theory as a *directed graph* G, consisting of a set of vertices $\mathbf{V}$, and a set of edges $\mathbf{E}$. Each edge $e_{ij} \in \mathbf{E}$ connects two vertices $v_i, v_j \in \mathbf{V}$. For consistency, we use the term *node* to refer to a vertex (Wikipedia article) in the graph, and *link* to refer to an edge (hyperlink) between such nodes.

In this model, each Wikipedia article is seen to represent a single *concept*, and the hyperlink structure relates these concepts to one another. In order to compute relatedness between two concepts v_i and v_j, we use spreading activation and rely on the fundamental principle of an associative network, namely that it connects nodes that are associated with one another via real-valued links denoting how strongly the objects are related. Since Wikipedia was not created as an associative network, but primarily as an online encyclopaedia, none of these weights exist, and we will have to deduce these (see *Fan-out constraint* in Section 4).

Links *into* pages are used, since this leads to better results (Witten and Milne, 2008). The Wikipedia graph structure is represented in an adjacency list structure, i.e. for each node v_i we store its list of neighbour nodes in a dictionary using v_i's id as key. This approach is preferred over an adjacency matrix structure, since most articles are linked to by only 34 articles on average, which would lead to a very sparse adjacency matrix structure.

4 Adapting Spreading Activation for Wikipedia's Hyperlink Structure

Each pulse in the Spreading Activation (SA) process consists of three stages: 1) pre-adjustment, 2) spreading, and 3) post-adjustment (Crestani, 1997). During pre- and post-adjustment, some form of activation decay is optionally applied to the active nodes. This serves both to avoid retention of activation from previous pulses, and, from a connectionist point of view, models 'loss of interest' when nodes are not continually activated.

Let $a_{i,\text{in}}$ denote the total energy input (activation) for node v_i, and $N(v_i)$ the set of v_i's neighbour nodes with incoming links to v_i. Also, let $a_{j,\text{out}}$ denote the output activation of a node v_j connected to node v_i, and let w_{ij} denote the weight of connection between node v_i and v_j. For a node v_i, we can then describe the pure model of spreading activation as follows:

$$a_{i,\text{in}} = \sum_{v_j \in N(v_i)} a_{j,\text{out}} w_{ij}. \qquad (1)$$

This pure model of SA has several significant problems, the most notable being that activation can saturate the entire network unless certain constraints are imposed, namely limiting how far activation can spread from the initially activated nodes (distance constraint), and limiting the effect of very highly-connected nodes (fan-out constraint) (Crestani, 1997). In the following three sections we discuss how these constraints were implemented in our model for SA.

Distance constraint

For every pulse in the spreading process, a node's activation value is multiplied by a global network decay parameter $0 < d < 1$. We therefore substitute w_{ij} in Equation 1 for $w_{ij}d$. This decays activation exponentially in the path length. For a path length of one, activation is decayed by d, for a path length of two, activation is decays by $dd = d^2$, etc. This penalises activation transfer over longer paths. We also include a maximum path length parameter $L_{p,\max}$ which limits how far activation can spread.

Fan-out constraint

As noted above, in an associative network, links have associated real-valued weights to denote the strength of association between the two nodes they connect (i.e. w_{ij} in Equation 1). These weights have to be estimated for the Wikipedia hyperlink graph, and for this purpose we propose the use of three **weighting schemes**:

In **pure Energy Distribution (ED)**, a node v_i's weight w is made inversely proportional to its in-degree (number of neighbours $N(v_i) \geq 1$ with incoming links to v_i[1]). Thus $\text{ED}(v_i, v_j) = w_{ij} = \frac{1}{|N(v_i)|}$. This reduces the effect of very connected nodes on the spreading process (constraint 2 above).

For instance, we consider a path connecting two nodes via a general article such as **`USA`** (connected to 322,000 articles) not nearly as indicative of a semantic relationship, as a path connecting them via a very **specific** concept, such as **`Hair Pin`** (only connected to 20 articles).

Inverse Link-Frequency (ILF) is inspired by the term-frequency inverse document-frequency (**tf-idf**) heuristic (Salton and McGill, 1983) in which a term's weight is reduced as it is contained in more documents in the corpus. It is based on the idea that the more a term appears in documents across the corpus, the less it can discriminate any one of those documents.

We define a node v_i's *link-frequency* as the number of nodes that v_i is connected to $|N(v_i)|$ divided by the number of possible nodes it could be connected to in the entire Wikipedia graph

[1] All orphan nodes are removed from the AN.

$|G|$, and therefore give the log-smoothed *inverse* link-frequency of node v_i as:

$$\mathrm{ILF}(v_i) \triangleq \log\left(\frac{|G|}{|N(v_i)|}\right) \geq 0 \qquad (2)$$

As noted above for *pure energy distribution*, we consider less connected nodes as more specific. If one node connects to another via a very **specific** node with a low in-degree, $\frac{|G|}{|N(v_i)|}$ is very large and $\mathrm{ILF}(v_i) > 1$, thus boosting that specific link's weight. This has the effect of 'boosting' paths (increasing their contribution) which contain nodes that are less connected, and therefore more meaningful in our model.

To evaluate the effect of this boosting effect described above, we also define a third normalised weighting scheme called the **Normalised Inverse Link-Frequency (NILF)**, $0 \leq \mathrm{NILF}(v_i) \leq 1$:

$$\mathrm{NILF}(v_i) \triangleq \frac{\mathrm{ILF}(v_i)}{\log|G|}. \qquad (3)$$

ILF reaches a maximum of $\log|G|$ when $|N(v_i)| = 1$ (see Equation 2). We therefore divide by $\log|G|$ to normalise its range to [0,1].

Threshold constraint

Finally, the above-mentioned constraints are enforced through the use of a threshold parameter $0 < T < 1$. Activation transfer to a next node ceases when a node's activation value drops below a certain threshold T.

5 Strategies for Interpreting Activations

After spreading has ceased, we are left with a vector of nodes and their respective values of activation (an **activation vector**). We wish to translate this activation vector into a score resembling strength of association or relatedness between the two initial nodes.

We approach this problem using two different approaches, the Target Activation Approach (TAA) and the Agglomerative Approach (AA). These approaches are based on two distinct hypotheses, namely: Relatedness between two nodes can be measured as either 1) the ratio of initial energy that reaches the target node, or 2) the amount of overlap between their individual activation vectors by spreading from both nodes individually.

Target Activation Approach (TAA)

To measure the relatedness between v_i and v_j, we set a_i to some initial value K_{init} (usually 1.0), and all node activations including $a_j = 0$. After the SA process has terminated, v_j is activated with some $a_{j,\mathrm{in}}$. Relatedness is computed as the ratio $\mathrm{sim}_{TAA}(v_i, v_j) \triangleq \frac{a_{j,\mathrm{in}}}{K_{\mathrm{init}}}$.

Agglomerative Approach (AA)

The second approach is called the Agglomerative Approach since we agglomerate all activations into one score resembling relatedness. After spreading has terminated, relatedness is computed as the amount of overlap between the individual nodes' activation vectors, using either the cosine similarity (AA-cos), or an adapted version of the information theory based WLM (Witten and Milne, 2008) measure.

Assume the same set of initial nodes v_i and v_j. Let $\mathbf{A}_k$ be the N-dimensional vector of real-valued activation values obtained by spreading over the N nodes in the graph from node v_k (called an **activation vector**). We use a_{kx} to denote the element at position x in $\mathbf{A}_k$. Furthermore, let $\mathbf{V}_k = \{v_{k1}, ..., v_{kM}\}$ denote the set of M nodes activated by spreading from v_k, i.e. the set of identifiers of nodes with non-zero activations in $\mathbf{A}_k$ after spreading has terminated (and therefore $M \leq N$).

We then define the **cosine Agglomerative Approach** (henceforth called **AA-cos**) as

$$\mathrm{sim}_{\mathrm{AA,cos}}(\mathbf{A}_i, \mathbf{A}_j) \triangleq \frac{\mathbf{A}_i \cdot \mathbf{A}_j}{||\mathbf{A}_i||||\mathbf{A}_j||} \qquad (4)$$

For our adaptation of the Wikipedia Link-based Measure (WLM) approach to spreading activation, we define the **WLM Agglomerative Approach** (henceforth called **AA-wlm**[2]) as

[2] *AA-wlm* is our adaptation of WLM (Witten and Milne, 2008) for SA, not to be confused with their method, which we simply call *WLM*.

$$\mathrm{sim}_{\mathrm{AA,wlm}}(\mathbf{V_i}, \mathbf{V_j}) \triangleq \frac{\log\big(\max(|\mathbf{V_i}|,|\mathbf{V_j}|)\big)-\log(|\mathbf{V_i}\cap\mathbf{V_j}|)}{\log(|G|)-\log(\min(|\mathbf{V_i}|,|\mathbf{V_j}|))} \quad (5)$$

with $|G|$ representing the number of nodes in the entire Wikipedia hyperlink graph. Note that the AA-wlm method does not take activations into account, while the AA-cos method does.

6 Spreading Activation Algorithm

Both the TAA and AA approaches described above rely on a function to spread activation from one node to all its neighbours, and iteratively to all their neighbours, subject to the constraints listed. TAA stops at this point and computes relatedness as the ratio of energy received to energy sent between the target and source node respectively. However, AA repeats the process from the target node and computes relatedness as some function (cosine or information theory based) of the two activation vectors, as given by Equation 4 and Equation 5.

We therefore define SPREAD_UNIDIR() as shown in Algorithm 1. Prior to spreading from some node v_i, its activation value a_i is set to some initial activation value K_{init} (usually 1.0). The activation vector $\mathbf{A}$ is a dynamic node-value-pair list, updated in-place. $\mathbf{P}$ is a dynamic list of nodes in the *path* to v_i to avoid cycles.

Algorithm 1 Pseudo code to spread activation depth-first from node v_i up to level $L_{p,\max}$, using global decay d, and threshold T, given an adjacency list graph structure G and a weighting scheme $\mathbf{W}$ such that $0 < w_{ij} \in \mathbf{W} < 1$.

```
Require: G, L_{p,max}, d, T
  function SPREAD_UNIDIR(v_i, A, P)
    if (v_i, a_i) ∉ A or a_i < T then                ▷ Threshold
      return
    end if
    Add v_i to P                                     ▷ To avoid cycles
    for v_j ∈ N(v_i) do                              ▷ Process neighbours
      if (v_j, a_j) ∉ A then
        a_j = 0
      end if
      if v_j ∉ P and |P| ≤ L_{p,max} then
        a*_j = a_j + a_i * w_ij * d
        Replace (v_j, a_j) ∈ A with (v_j, a*_j)
        SPREAD_UNIDIR(v_j, A, P)
      end if
    end for
    return
  end function
```

7 Parameter Optimisation: Inter-concept Similarity

The model for SA as introduced in this paper relies on several important parameters, namely the spreading strategy (TAA, AA-cos, or AA-wlm), weighting scheme (pure ED, ILF, and NILF), maximum path length $L_{p,\max}$, network decay d, and threshold T. These parameters have a large influence on the accuracy of the proposed technique, and therefore need to be optimised.

Experimental Method

In order to compare our method with results reported by Gabrilovich and Markovitch (2007) and Witten and Milne (2008), we followed the same approach by randomly selecting 50 word-pairs from the WordSimilarity-353 dataset (Gabrilovich, 2002) and correlating our method's scores with the human-assigned scores. To reduce the possibility of overestimating the performance of our technique on a sample set that happens to be favourable to our technique, we furthermore implemented a technique of **repeated holdout** (Witten and Frank, 2005):

Given a sample test set of N pairs of words with human-assigned ratings of relatedness, randomly divide this set into k parts of roughly equal size[3]. Hold out one part of the data and iteratively evaluate the performance of the algorithm on the remaining $k-1$ parts until all k parts have been held out once. Finally, average the algorithm's performance over all k runs into one score resembling the performance for that set of parameters.

Since there are five parameters (spreading strategy, weighting scheme, path length, network decay, and threshold), a **grid search** was implemented by holding three of the five parameters constant, and evaluating combinations of decay and threshold by stepping over the possible parameter space using some step size. A coarse-grained grid search was first conducted with step

[3] k was chosen as 5.

Table 1: Spreading results by spreading strategy (TAA=Target Activation Approach, AA=Agglomerative Approach, $L_{p,\max}$ = maximum path length used, ED=energy distribution only, ILF=Inverse Link Frequency, NILF=normalised ILF.) Best results in bold.

Strategy	ρ_{max}	Parameters
TAA	0.56	ED, $L_{p,\max}$=3, d=0.6, T=0.001
AA-wlm	0.60	NILF, $L_{p,\max}$=3, d=0.1, T=10^{-6}
AA-cos	0.70	ILF, $L_{p,\max}$=3, d=0.5, T=0.1

size of 0.1 over d and a logarithmic scale over T, thus $T = \{0, 0.1, 0.01, 0.001, ..., 10^{-9}\}$. The best values for d and T were then chosen to conduct a finer-grained grid search.

Influence of the different Parameters

The **spreading strategy** determines how activations resulting from the spreading process are converted into scores of relatedness or similarity between two nodes. Table 1 summarises the best results obtained for each of the three strategies, with the specific set of parameters that were used in each run.

Results are better using the AA ($\rho_{max} = 0.70$ for AA-cos) than using the TAA ($\rho_{max} = 0.56$). Secondly, the AA-cos spreading strategy significantly outperforms the AA-wlm strategy over this sample set ($\rho_{\max,\text{wlm}} = 0.60$ vs $\rho_{\max,\cos} = 0.70$). These results compare favourably to similar inter-concept results reported for WLM (Witten and Milne, 2008) ($\rho = 0.69$) and ESA (Gabrilovich and Markovitch, 2007) ($\rho = 0.75$).

Maximum path length $L_{p,\max}$ is related to how far one node can spread its activation in the network. We extend the first-order link model used by WLM, by approaching the link structure as an associative network and by using spreading activation.

To evaluate if this is a useful approach, tests were conducted by using maximum path lengths of one, two, and three. Table 2 summarises the results for this experiment. Increasing path length from one to two hops increases performance from $\rho_{max} = 0.47$ to $\rho_{max} = 0.66$. Moreover, increasing $L_{p,\max}$ from two to three hops furthermore increases performance to $\rho_{max} = 0.70$.

Table 2: Spreading results by maximum path length $L_{p,\max}$. Best results in bold.

$L_{p,\max}$	ρ_{max}	Parameters
1	0.47	TAA, ED/ILF/NILF
2	0.66	AA-cos, ILF, d=0.4, T=0.1
3	0.70	AA-cos, ILF, d=0.5, T=0.1

Table 3: Spreading results by weighting scheme w. Best results in bold.

w	ρ_{max}	Parameters
NILF	0.63	AA-cos, $L_{p,\max} = 3$, d=0.9, T=0.01
ED	0.64	AA-cos, $L_{p,\max} = 3$, d=0.9, T=0.01
ILF	0.70	AA-cos, $L_{p,\max} = 3$, d=0.5, T=0.1

In an associative network, each link has a real-valued weight denoting the *strength of association* between the two nodes it connects. The derived Wikipedia hyperlink graph lacks these weights. We therefore proposed three new **weighting schemes** (pure ED, ILF, and NILF) to estimate these weights.

Table 3 summarises the best performances using the different weighting schemes. ILF outperforms both ED and NILF. Furthermore, both ED and NILF perform best using higher decay values (both 0.9) and lower threshold values (both 0.01), compared to ILF (0.5 and 0.1 respectively for d and T). We attribute this observation to the boosting effect of the ILF weighting scheme for less connected nodes, and offer the following explanation:

Recall from the section on ILF that in our model, strongly connected nodes are viewed as more general, and nodes with low in-degrees are seen as very **specific** concepts. We argued that a path connecting two concepts via these more specific concepts are more indicative of a stronger semantic relationship than through some very general concept. In the ILF weighting scheme, paths containing these less connected nodes are automatically boosted to be more im-

portant. Therefore, by not boosting less meaningful paths, a lower decay and higher threshold effectively limits the amount of non-important nodes that are activated, since their activations are more quickly decayed, whilst at the same time requiring a higher threshold to continue spreading. Boosting more important nodes can therefore lead to activation vectors which capture the semantic context of the source nodes more accurately, leading to higher performance.

8 Computing document similarity

To compute document similarity, we first extract key representative Wikipedia concepts from a document to produce **document concept vectors**[4]. This process is known as *wikification* (Csomai and Mihalcea, 2008), and we used an implementation of Milne and Witten (2008). This produces document concept vectors of the form $\mathbf{V_i} = \{(id_1, w_1), (id_2, w_2), ...\}$ with id_i some Wikipedia article identifier and w_i a weight denoting how strongly the concept relates to the current document. We next present two algorithms, MAXSIM and WIKISPREAD, for computing document similarity, and test these over the Lee (2005) document similarity dataset, a set of 50 documents between 51 and 126 words each, with the averaged gold standard similarity ratings produced by 83 test subjects (see (Lee et al., 2005)).

The first metric we propose is called **MAXSIM** (see Algorithm 2) and is based on the idea of measuring document similarity by pairing up each Wikipedia concept in one document's concept vector with its most similar concept in the other document. We average those similarities to produce an inter-document similarity score, weighted by how strongly each concept is seen to represent a document ($0 < p_i < 1$). The contribution of a concept is further weighted by its ILF score, so that more specific concepts contribute more to final relatedness.

The second document similarity metric we propose is called the **WIKISPREAD** method and is a natural extension of the inter-concept spreading activation work introduced in the previous section. We view a document concept vector as a cluster of concepts, and build a single *document activation vector* (see Algorithm 3) – i.e. a vector of article ids and their respective activations – for each document, by iteratively spreading from each concept in the document concept vector. Finally, similarity is computed using either the AA-cos or AA-wlm methods given by Equation 4 and Equation 5 respectively.

Knowledge-based approaches such as the Wikipedia-based methods can capture more complex lexical and semantic relationships than

[4]Vectors of Wikipedia topics (concepts) and how strongly they are seen to relate to the current document.

Algorithm 2 Pseudo code for the MaxSim algorithm for computing inter-document similarity. v_i is a Wikipedia concept and $0 < p_i < 1$ how strongly it relates to the current document.

```
Require: ILF lookup function
  function MAXSIM(V1, V2)
    num=0
    den=0
    for (v_i, p_i) ∈ V1 do
      s_k = 0                          ▷ s_k = max_j sim(v_i, v_j)
      for v_j ∈ V2 do                  ▷ Find most related topic
        s_j = sim(v_i, v_j)
        if s_j > s_k then
          v_k = v_j ▷ Topic in V2 most related to v_i
          s_k = s_j
        end if
      end for
      num += s_k p_i ILF(v_k)
      den += ILF(v_k)
    end for
    return num / den
  end function
```

Algorithm 3 Pseudo code for the WikiSpread algorithm for computing inter-document similarity. $K_{\text{init}} = 1.0$.

```
  function WIKISPREAD(V1, V2)
    A1 = ∅                             ▷ Dynamic activation vectors.
    A2 = ∅
    for (v_i, p_i) ∈ V1 do             ▷ Document 1
      a_i = K_init · p_i               ▷ Update a_i ∝ p_i
      Add (v_i, a_i) to A1
      SPREAD_UNIDIR(v_i, A1, ∅)
    end for
    for (v_j, p_j) ∈ V2 do             ▷ Document 2
      a_j = K_init · p_j
      Add (v_j, a_j) to A2
      SPREAD_UNIDIR(v_j, A2, ∅)
    end for
    Compute similarity using AA-cos or AA-wlm
  end function
```

Table 4: Summary of final document similarity correlations over the Lee & Pincombe document similarity dataset. ESA score from Gabrilovich and Markovitch (2007).

	Pearson ρ
Cosine VSM (with tf-idf) only	0.56
MaxSim method	0.68
WikiSpread method	0.62
ESA	0.72
Combined (Cosine + MaxSim)	**0.72**

keyword-matching approaches, however, nothing can be said about concepts not adequately represented in the underlying knowledge base (Wikipedia). We therefore hypothesise that combining the two approaches will lead to more robust document similarity performance. Therefore, the final document similarity metric we evaluate (**COMBINED**) is a linear combination of the best-performing Wikipedia-based methods described above, and the well-known Vector Space Model (VSM) with cosine similarity and tf-idf (Salton and McGill, 1983).

Results

The results obtained on the Lee (2005) document similarity dataset using the three document similarity metrics (MAXSIM, WIKISPREAD, and COMBINED) are summarised in Table 4. Of the two Wikipedia-only methods, the MaxSim method achieves the best correlation score of $\rho = 0.68$. By combining the standard cosine VSM with tf-idf with the MaxSim metric in the ratio λ and $(1 - \lambda)$ for $0 < \lambda < 1$, and performing a parameter sweep over λ, we can weight the contributions made by the individual methods and observe the effect this has on final performance. The results are shown in Fig 1. Note that both methods contribute equally ($\lambda = 0.5$) to the final best correlation score of $\rho = 0.72$. This suggests that selective knowledge-based augmentation of simple VSM methods can lead to more accurate document similarity performance.

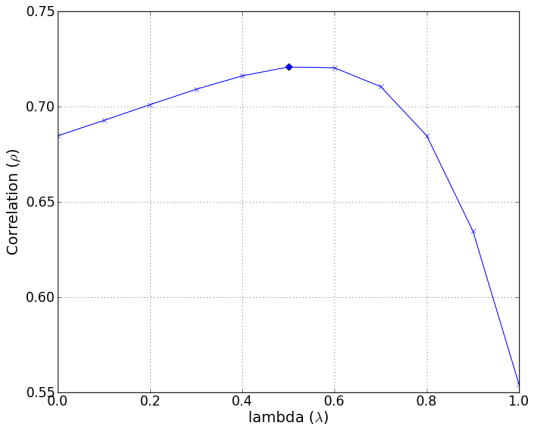

Figure 1: Parameter sweep over λ showing contributions from cosine (λ) and Wikipedia-based MAXSIM method $(1 - \lambda)$ to the final performance over the Lee (2005) dataset.

9 Conclusion

In this paper, the problem of computing conceptual similarity between concepts and documents are approached by spreading activation over Wikipedia's hyperlink graph. New strategies are required to infer weights of association between articles, and for this we introduce and test three new weighting schemes and find our Inverse Link-Frequency (ILF) to give best results. Strategies are also required for translating resulting activations into scores of relatedness, and for this we propose and test three new strategies, and find that our cosine Agglomerative Approach gives best results. For computing document similarity, we propose and test two new methods using only Wikipedia. Finally, we show that using our best Wikipedia-based method to augment the cosine VSM method using tf-idf, leads to the best results. The final result of $\rho = 0.72$ is equal to that reported for ESA (Gabrilovich and Markovitch, 2007), while requiring less than 10% of the Wikipedia database required for ESA. Table 4 summarises the document-similarity results.

Acknowledgements

We thank Michael D. Lee for his document similarity data and MIH Holdings Ltd. for financially supporting this research.

References

Berger, Helmut, Michael Dittenbach, and Dieter Merkl. 2004. An adaptive information retrieval system based on associative networks. *APCCM '04: Proceedings of the first Asian-Pacific conference on Conceptual Modelling*, pages 27–36.

Budanitsky, A. and G. Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2. Citeseer.

Collins, A.M. and E.F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407–428.

Crestani, F. 1997. Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review*, 11(6):453–482.

Csomai, A. and R. Mihalcea. 2008. Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems*, 23(5):34–41.

Gabrilovich, E. and S. Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12.

Gabrilovich, E. 2002. The WordSimilarity-353 Test Collection. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*.

Lee, M.D., B. Pincombe, and M. Welsh. 2005. A Comparison of Machine Measures of Text Document Similarity with Human Judgments. In *27th Annual Meeting of the Cognitive Science Society (CogSci2005)*, pages 1254–1259.

Lyman, P. and H.R. Varian. 2003. How much information? `http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/index.htm`. Accessed: May, 2010.

Metzler, Donald and W. Bruce Croft. 2006. Beyond bags of words: Modeling implicit user preferences in information retrieval. *AAAI'06: Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1646–1649.

Milne, David and Ian H. Witten. 2008. Learning to link with wikipedia. *CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pages 509–518.

Preece, SE. 1982. *Spreading Activation Network Model for Information Retrieval.* Ph.D. thesis.

Salton, G. and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill New York.

Tran, T., P. Cimiano, S. Rudolph, and R. Studer. 2008. Ontology-based Interpretation of Keywords for Semantic Search. *The Semantic Web*, pages 523–536.

Witten, I.H. and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Witten, I.H. and D. Milne. 2008. An Effective, Low-Cost Measure of Semantic Relatedness Obtained From Wikipedia Links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30.

Yeh, E., D. Ramage, C.D. Manning, E. Agirre, and A. Soroa. 2009. WikiWalk: Random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 41–49. Association for Computational Linguistics.