

# Inference of Relevancy:

## An Introduction to Vector Spread Activation

It is a well established fact that major search engines are exploiting various types of inter-connectivity data to enhance search query relevance. In particular; formulas such as [PageRank\(tm\)](#) make objective valuations of hyper-documents via citations information alone. Each search index has its own scheme which similarly falls into the category of link popularity analysis. This however is not the final extent to which link graph data can assist in document retrieval, among the many unique and innovative approaches for leveraging link information within an IR system; associative relevancy inference techniques are among the most interesting. In this paper we will be focusing on **Vector Spread Activation** in theory and practice.

For purposes of illustration we will begin with a hypothetical web page: (A) Let's say that (A) is a document within a well respected website about shovels, we'll call it *All Manner of Shovels dot com*.

Our page (A) specifically deals with gardening shovels and nothing else. Furthermore (A) offers its visitors a valuable extension to its own content by providing a link to an external webpage that deals exclusively with hand rakes, we'll call this page (B). Page (B) bears the great distinction of being cited (through hyper-links) by a variety of gardening and gardening tool webpages, however (B) is not linked to by any other webpages that deal with hand rakes specifically; as any such webpage would probably be a competitor to (B).

Now in a classical IR system any queries for "hand rake" would retrieve our page (B). Depending on the similarity thresholds in place any queries for either "hand" or "rake" alone could possibly retrieve (B) as well.

It could also be said that in a classical IR system a query for "garden trowel" would NOT return page (B), however it would return some of the pages that link to (B), one such page in particular (C), is an excellent resource for garden trowels and would definitely be returned for such a query.

The problem with this approach is that (B) and (C) compliment each other in the sense that (B)'s hand rakes are a logical companion of (C)'s trowels in terms of real world application. Therefore it stands to reason that someone searching for garden trowels would find information for hand rakes to be both valuable and relevant for their query.

The solution to this problem is Vector Spread Activation or VSA. VSA states that if a webpage (p) is linked to by many documents(p1...p2) which are relevant to a query (q), yet possesses a low similarity itself for the query (q), then it makes sense to add a portion of the similarities of its parents (p1...p2) to increase the probability that (p) will also be retrieved for that query (q).

The interesting feature of VSA is that it allows for documents which do not contain any part of a query to be returned for that query, furthermore a document in this scenario could rank higher than documents that contain the actual search query terms.

Vector Spread Activation is calculated as follows:

For a query (q) there exists a regular similarity  $sim(q,p1)$  between (q) and a given webpage (p1) (This is usually calculated using a variation of Vector Space Model Term Weighting ). Next we introduce a function

$c(x,y)$  which equals 1 if a page (px) links to a page (py) , otherwise  $c(x,y)$  equals 0.

A VSA ranking score of a document for a given query  $rs(q, p1)$  is computed by:

$$rs(q,p1) = sim(q,p1) + \alpha \sum_{p2} c(p2,p1) * sim(q,p2)$$

Where:

- $\alpha$  = The amount of a citing page's similarity score that will be passed to the cited page.
- $p2$  = The pages which cite  $p1$ .
- $Sim(q,p2)$  = The similarity scores between our citing pages  $p2$  and our query (q).
- $rs$  = The similarity score of 1 now augmented by a portion of the similarity scores of the pages which cite it.

In practice an initial set of relevant documents are collected for a search query; the formula is then be applied to the pages which are linked to by the candidate documents. Those which are linked to by many of the candidate documents are likely to be included in the search results as their relevancy has been inferred by the similarity of the documents which point to them.