ABSTRACT

**Phrase Based Indexing and its use in spam detection**

Phrase Based indexing is a process used to index or retrieve information by utilizing phrases to list and search documents on the internet. It is an extension of the postings list concept.

Phrases are analysed to decide if they are “valid” or “good” phrases. The relevance of a phrase is judged by how frequently it is used and also relation between phrases is analysed (certain phrases are mentioned together with other phrases in the same document). A spam document contains an excessive number of related phrases.

I will speak briefly about how phrase based indexing works and also about a patent developed at Google that identifies phrases and accordingly indexes documents to identify spam documents.

URL for patent: http://www.google.com/patents/US7603345