(and user-unfriendly) Boolean logic is employed. Web search engines can be slow, although faster search engines are being developed, and matching is often poor (quantity does not necessarily indicate quality) as Search Engines often employ simple keyword pattern matching that takes no account of relevance. Search Engines often simply return the document with the greatest number of keyword occurrences. A methodology to process documents unsupervised, handle paraphrasing of documents, to focus retrieval by minimising the search space and to automatically calculate the document similarity from statistics available in the text corpus is desired. Document may be clustered according to the user's requirements (clustered 'on the fly') and then employ category-specific finer-grained matching techniques.

Word categorisation (encompassing both unsupervised clustering and supervised classification) enables the words to be associated or grouped according to their meaning to produce a thesaurus. In this paper we focus solely on word clustering as this approach is unsupervised. Clustering does not require pre-generated human classifications to train the algorithm and is therefore less subjective and more automated as it learns from text corpus knowledge only. Word clustering can also overcome the *Vocabulary Problem* cited by Chen et al. [2]. They posit that through the diversity of expertise and background of authors and the polysemy of language, there are many ways to describe the same concept; there are many synonyms. In fact, Stetina et al. [20] postulate that polysemous words occur most frequently in text corpora even though most words in a dictionary are monosemous. Humans are able to intuitively cluster documents from imputed similarity. They overcome the differing vocabularies of authors and the inherent synonymy and polysemy of language. A computerised system must be able to match this ability. For computerised document similarity calculation, an underlying hierarchical synonym clustering is required to enable differing vocabularies to be accommodated. The distances in the hierarchy may be used for word similarity estimation and to score document similarity, thus allowing paraphrased documents to be awarded high similarity scores as their contained words fall into identical or neighbouring synonym clusters. Human generated thesauri are too general; they encompass all senses of words even though many are redundant for a particular domain. They are expensive with respect to construction time particularly if a single human knowledge engineer generates the hierarchy. If multiple experts are consulted then it is very difficult to obtain a single unified hierarchy. Human thesauri also omit certain senses and subdivide others where there is little distinction; they are rather subjective. Automatic methods can be trained generally or domain specifically as required. The hierarchy allows us to focus searching to cohesive clusters therefore minimising the search space for each query. In this paper we analyse current word categorisation approaches and describe and evaluate our method with respect to the current implementations. We compare our TreeGCS clustering method [7], [6] and sections 3.2 and 3.3 to the Self-Organising Map (SOM) [11] method and then compare