With the development of the Internet the well-known classic **Information Retrieval Problem**: *given a set of documents and a query, determine the subset of documents relevant to the query*, gained its modern counterpart in the form of the **Web Search Problem** described by [Selberg, 99]: *find the set of documents on the Web relevant to a given user query*. A broad range of so-called **web search engines** has emerged to deal with the latter task, Google and AllTheWeb being two examples of general-purpose services of this type. Available are also search engines that help the users locate very specific resources, such as CiteSeer which finds scientific papers and Froogle, which is a web shopping search engine. Practical and indespensable as all these services are, their functioning can still be improved.

1.1 Motivation

Low precision searches

The vast majority of publically available search engines adopt a so-called **query-list paradigm**, whereby in respose to a user's query the search engine returns a linear ranking of documents matching that query. The higher on the list, the more relevant to the query the document is supposed to be. While this approach works efficiently for well-defined narrow queries, when the query is too general, the user will have to sift through a large number of irrelevant documents in order to identify the ones they were interested in. This kind of situation is commonly referred to as a **low precision search**.

As shown in, more than 60% of web queries consist of one or two words, which inevitably leads to a large number of low precision searches. Several methods of dealing with the results of such searches have been proposed. One method is pruning of the result list ranging from simple duplicate removal to advanced The most common approach is **relevance feedback**, whereby the search engine asissts the user in finding additional key words that would make the query more precise and reduce the number of returned documents. An alternative and increasingly popular method is search results clustering.

Search Results Clustering

Search Results Clustering is a process of organising the document references returned by a search engine into a number of thematic categories. In this setting, for a query "Sheffield", for example, the user would be presented with such topical groups as "University of Sheffield", "Sheffield United", "Botanical gardens", "BBC Radio Sheffield" etc. In this way, the users gain insight into the structure of sub-topics covered by the results.

The idea of web search results clustering was first introduced in the Scatter/Gather system, which was based

1.2 The goal and scope of work

The aim of this project is to compare how different dimensionality reduction techniques will perform as part of the description-comes-first search results clustering algorithm. In particular, three techniques will be evaluated: Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF) and Local Non-negative Matrix Factorization. Thus, a separate version of Lingo should be designed and implemented for each of these techniques.