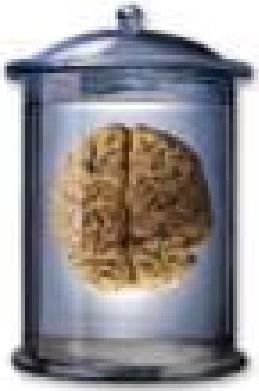
Machine learning techniques for document selection



vs.

Google™

Leon Derczynski - Supervised by Dr Amanda Sharkey - 2006

Search engines intend to show relevant documents

Keyword frequencies and document linking are used to choose content

Excellent!

This abstract relates to a document about low-price movies

Rubbish

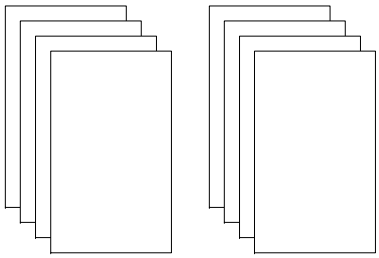
This document contains the words "cheap film", but is not useful

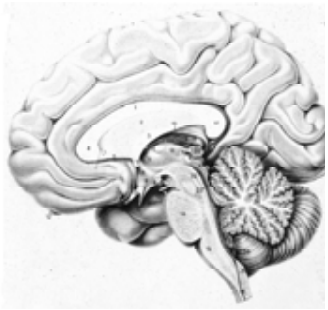
- Little human feedback is gathered on what makes a document relevant; it's mainly automated.

- The algorithms that decide relevancy are extremely complex and need to built from scratch. In 2003, Google used over 120 independent variables to sort results.

Is it possible to teach a system how to identify relevant documents without defining any explicit rules?

Pre-built sets of queries and relevant documents



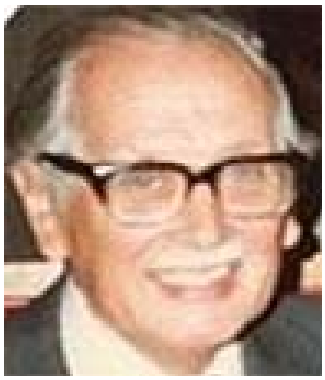


To teach a system how to distinguish relevant documents from irrelevant, a large amount of training data is required.

A wide range of documents and queries are needed to give a realistic model.

Early work in indexing documents – dating back to the 1960s – provides collections of sample queries, matched up to relevant document content.

Cyril Cleverdon pioneered work on organising information, and creating indexes. He led creation of a 1400-strong set of aerospace documents, accompanied by hundreds of natural language queries.



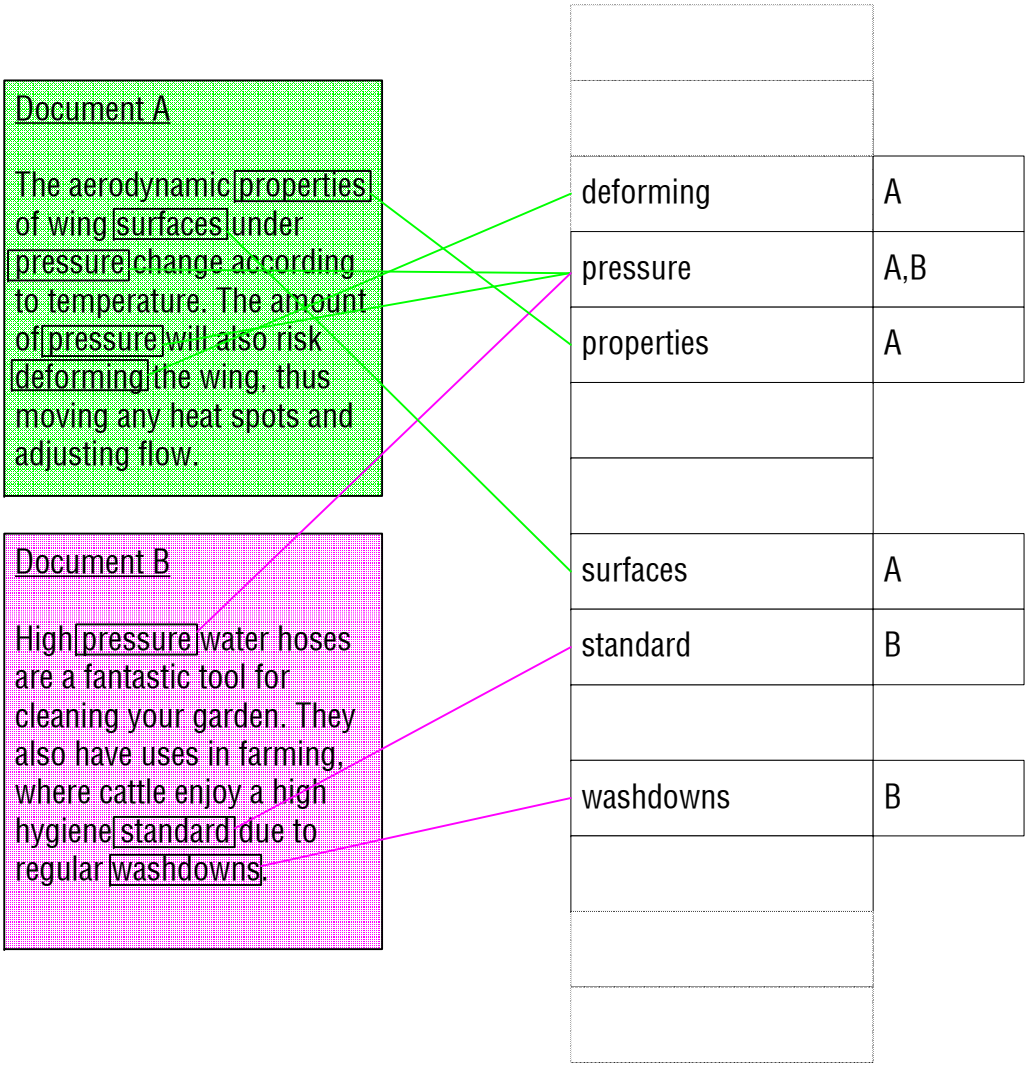
A list of matching documents was also manually created for each query.

This set of documents, queries and relevance judgements were known as the **Cranfield Collection**

Indexing documents

Searching all documents for a given query is a very time consuming process. Documents can be indexed according to the words they contain.

This shrinks search space considerably.



This allows documents containing keywords to be rapidly identified – only one lookup needs to be performed for each word in the query!

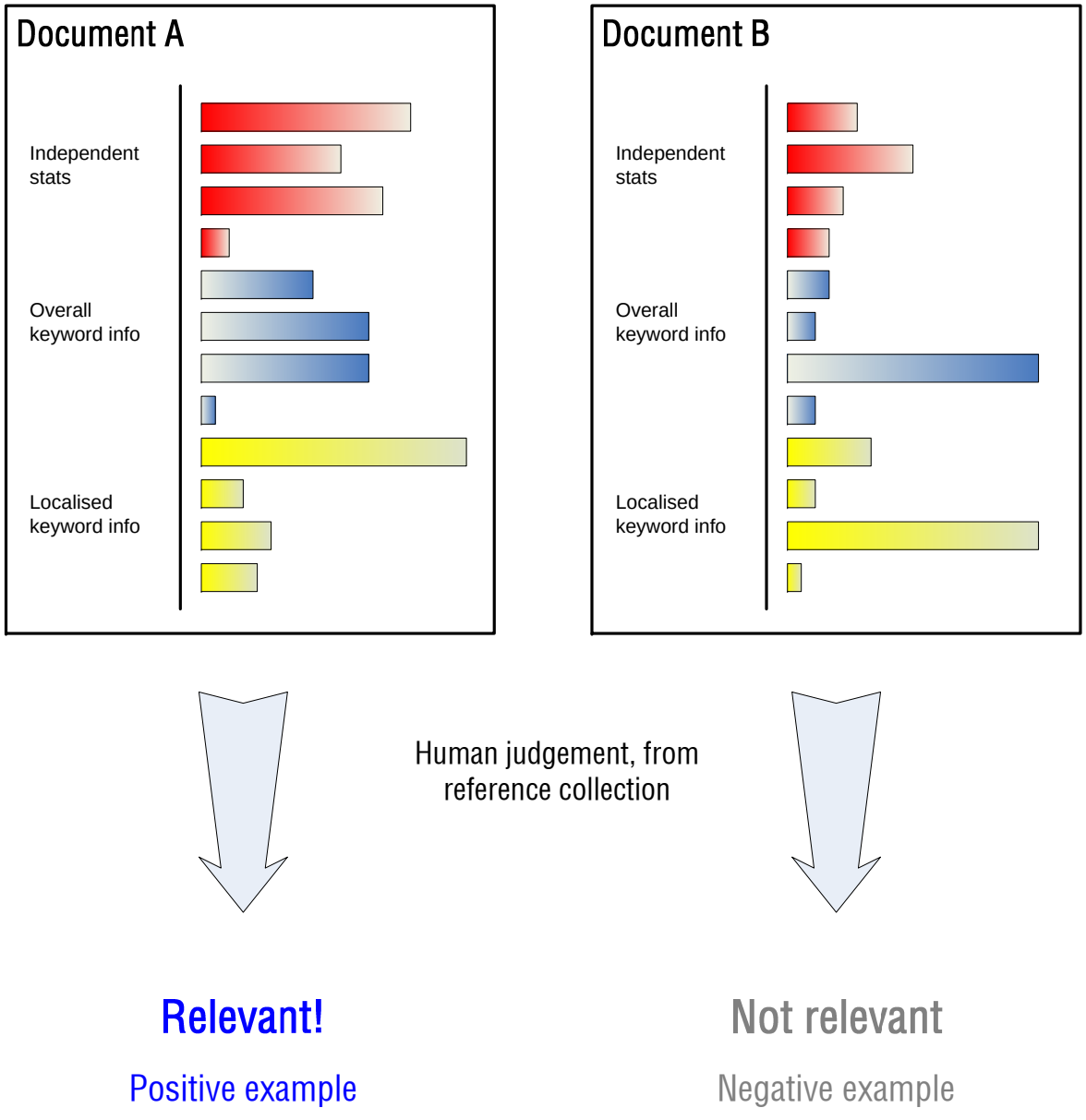
Identify document features

A set of statistics can be used to describe a document. They can be about the document itself, or about a particular word in the document. These numeric descriptions then become training examples for a machine learning algorithm.

For example, two documents can be assessed based on a query such as:

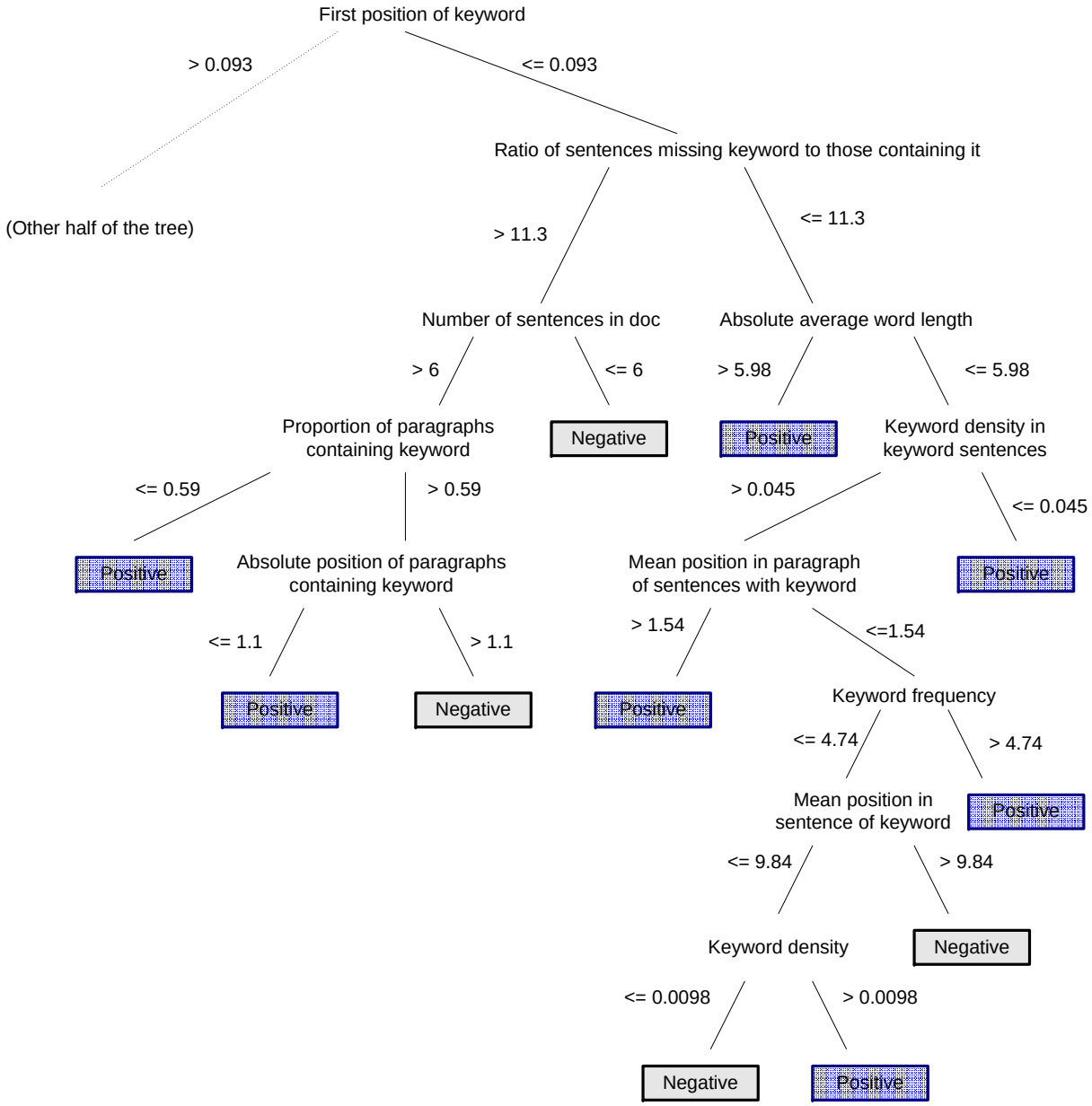
```
“what chemical kinetic system is applicable to
hypersonic aerodynamic Problems”
```

A set of statistics describing each document relative to the query can then be derived.



Training machine learning algorithms

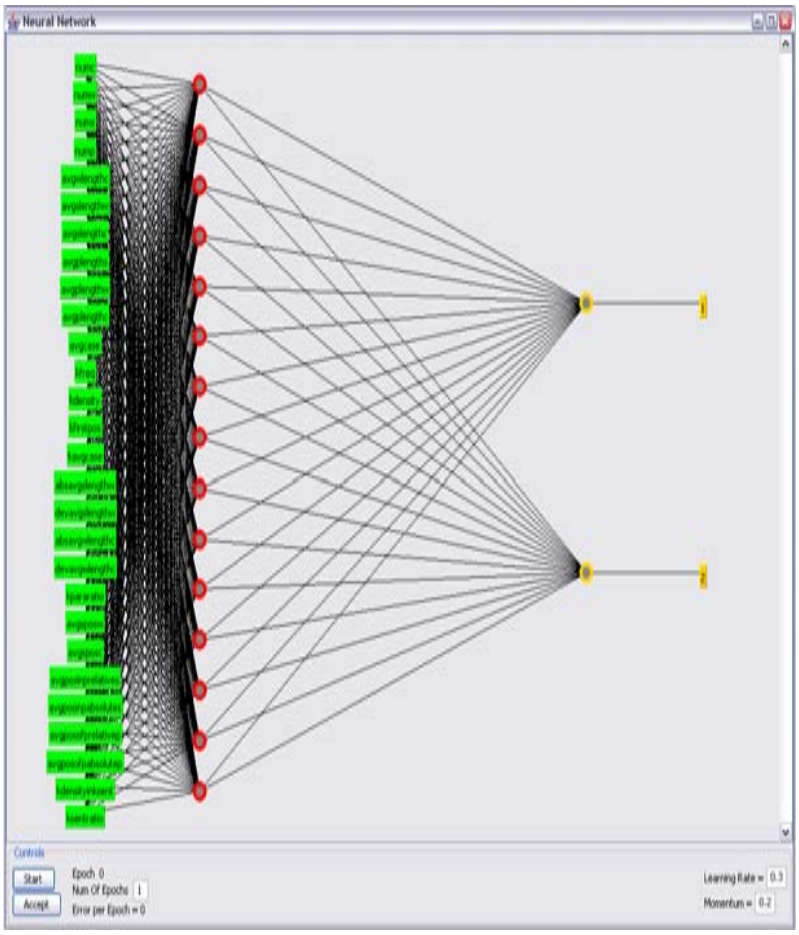
Decision trees are acyclic graphs that have a decision at each branch, based on an attribute of an example, and end at leaves which classify a document as relevant or not relevant.



A C4.5 Decision Tree, produced in an effort to emulate the decisions of the Cranfield judges.

The full version of this tree attained an 80.4% accuracy rate.

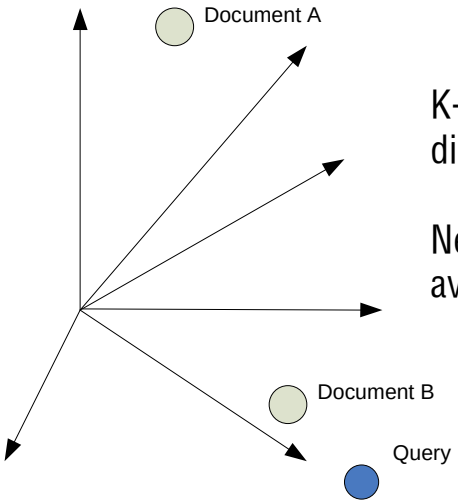
Neural nets



Neural nets have a set of nodes, each of which has various weights assigned to inputs. These are coupled with attributes, and when a certain internal value is reached, the output value changes.

Backpropagation is used to help converge on a net that solves the problem.

K-Nearest Neighbour



K-Nearest Neighbour plots all training data as points in multi-dimensional space, with one dimension for each attribute.

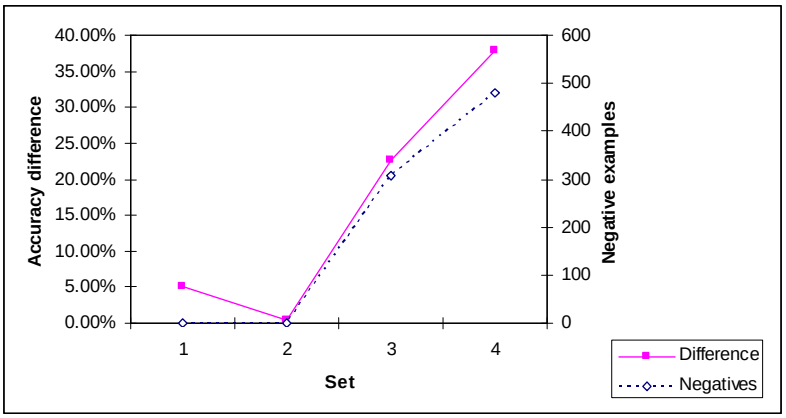
New examples are classified by working out the weighted average classification of the k nearest training examples.

Findings

The task is possible, with all algorithms managing to learn to identify a good amount of relevant documents.

Algorithm	Base accuracy	Average trained accuracy	Average improvement
Naïve Bayes	50.00%	75.39%	50.79%
C4.5 decision tree	50.00%	81.72%	63.44%
K* lazy	50.00%	67.26%	34.51%
N	50.00%	84.07%	68.14%

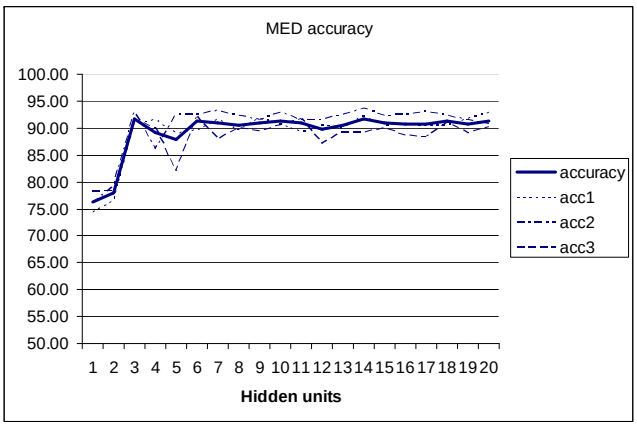
Neural nets showed the biggest overall improvement.



Both negative and positive examples are needed to train an algorithm effectively

---

Neural net accuracy improves as more units are added



Not every document suggested as relevant by human judges could be matched by the system. Sometimes, words were used that did not occur in the document.

Adding synonym lookup or a thesaurus should help.