Automatic identification of the most important elements in an XML collection

CSIRO ICT Centre
Alex Krumpholz

RSCS, Australian National University
Amir Hadad
Tom Gedeon

Uni. of Applied Science, Technikum Wien
Nina Studeny

Funnelback and RSCS, ANU
David Hawking
## INEX Collections

<table>
<thead>
<tr>
<th>Collection</th>
<th>No. Docs</th>
<th>No. distinct tags</th>
<th>No. tags judged relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE</td>
<td>16k</td>
<td>178</td>
<td>122</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>660k</td>
<td>1257</td>
<td>72</td>
</tr>
</tbody>
</table>
4.2.6 Forward Index

The forward index is actually already partially sorted. It is stored in a number of barrels (we used 64). Each barrel holds a range of wordID’s. If a document contains words that fall into a particular barrel, the docID is recorded into the barrel, followed by a list of wordID’s with hitlists which correspond to those words. This scheme requires slightly more storage because of duplicated docIDs but the difference is very small for a reasonable number of buckets and complexity in the final indexing phase done by the sorter. For wordID’s, we store each wordID as a relative difference from

---

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy
Automating

- Lots of features
- Fuzzy c-Means Clustering (Bezdek 1981) into 'relevant' and 'irrelevant'. Choose best cut-point.
- Measure alignment of clusters with those based on judgments. F-measure.
- Look at which features most closely align
- Eight features in common across Wikipedia and IEEE