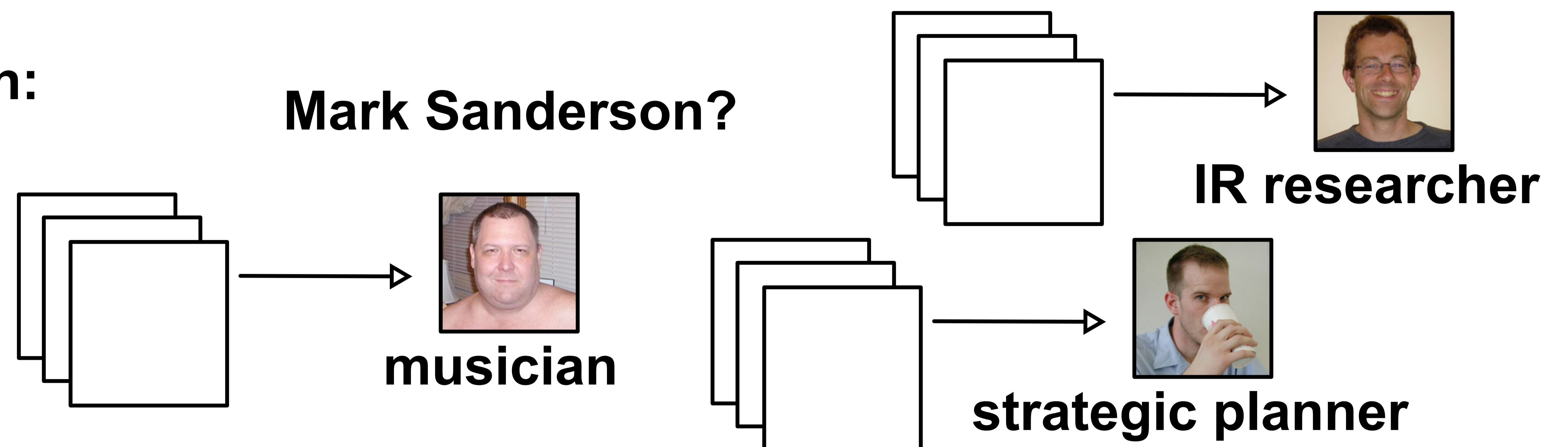


A Comparison of Retrieval-Based Hierarchical Clustering Approaches to Person Name Disambiguation

task

Person Name Disambiguation:

Group search results that refer to the same individual

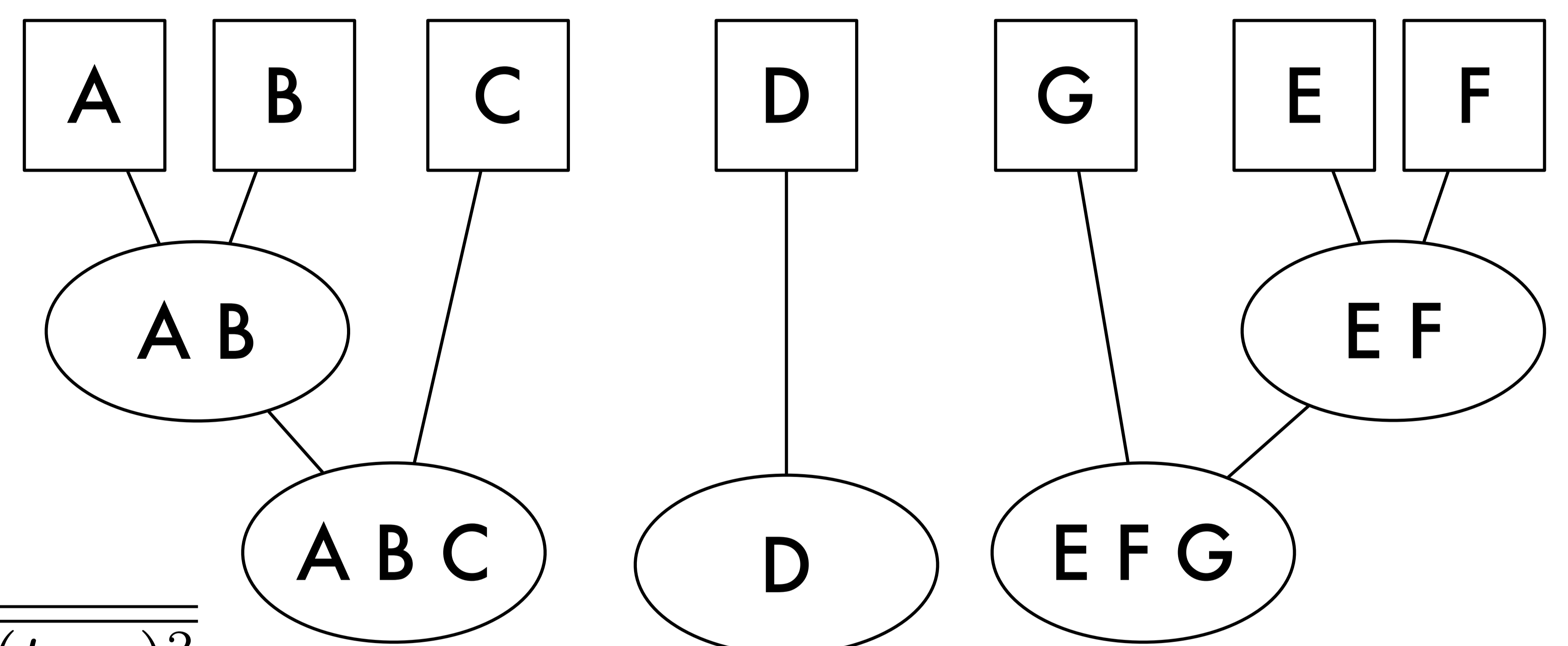


approach

Agglomerative Hierarchical Clustering:

Start with singleton clusters and group the most similar documents. Repeat this for document sets whenever a minimum similarity threshold is met. Uses cosine similarity:

$$sim(c_i, c_j) = \frac{\sum_{t \in c_i \cap c_j} w(t, c_i) \cdot w(t, c_j)}{\sqrt{\sum_{t' \in c_i} w(t', c_i)^2} \cdot \sqrt{\sum_{t' \in c_j} w(t', c_j)^2}}$$



results

Metrics: B³-Precision, B³-Recall, B³-F-score [1]

Collections: WePS-1 (dev/test) [2], WePS-2 (test) [3]

cluster representation	B ³ -P	B ³ -R	B ³ -F
maximum	0.937	0.493	0.576
minimum	0.860	0.724	0.754
centroid	0.886	0.666	0.720

term window	B ³ -P	B ³ -R	B ³ -F
50 words from name	0.821	0.787	0.772
no window	0.822	0.777	0.770

term frequencies	B ³ -P	B ³ -R	B ³ -F
normalized	0.822	0.777	0.770
absolute	0.579	0.893	0.651

collection frequencies	B ³ -P	B ³ -R	B ³ -F
local	0.769	0.692	0.679
collection	0.822	0.777	0.770
web crawl	0.608	0.883	0.672
wikipedia	0.495	0.934	0.581

term representation	B ³ -P	B ³ -R	B ³ -F
surface	0.860	0.724	0.754
stemmed	0.822	0.777	0.770
stemmed + hypernyms	0.860	0.706	0.739

conclusions

Efficient clustering methods using frequency-based statistics lead to high performance. Small variations in the definition of the similarity function have a substantial impact on performance.

[1] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*
 [2] J. Artiles, J. Gonzalo, and S. Sekine. The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. *SemEval 2007*
 [3] J. Artiles, J. Gonzalo, and S. Sekine. WePS 2 evaluation campaign: overview of the web people search clustering task. *WePS 2009, WWW 2009*