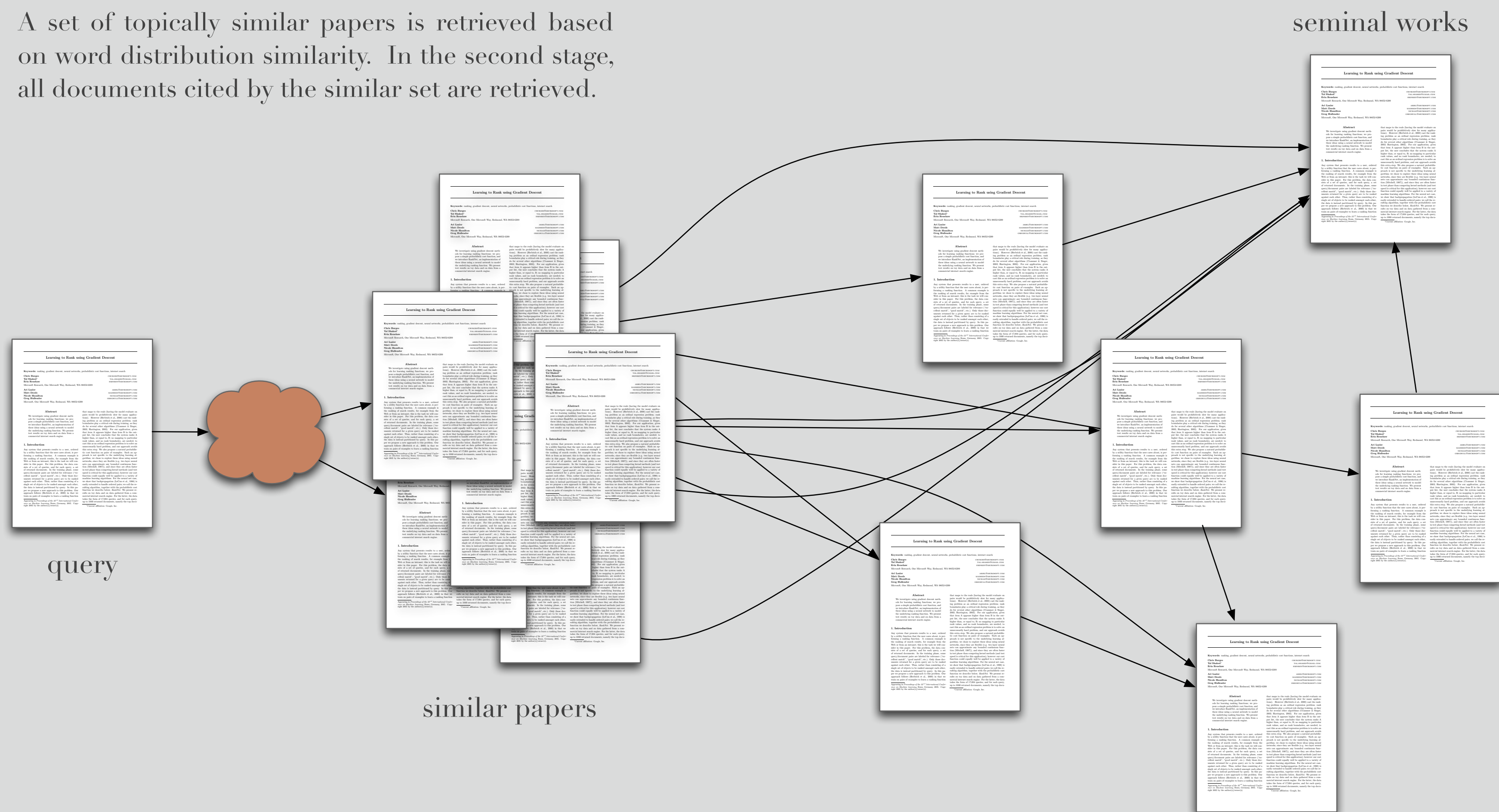


Recommending Citations for Academic Papers

Trevor Strohman, W. Bruce Croft, David Jensen

University of Massachusetts Amherst

A set of topically similar papers is retrieved based on word distribution similarity. In the second stage, all documents cited by the similar set are retrieved.



Problem

Researchers need to find relevant related work
 Scientific literature is vast and difficult to quantify
 Not always clear what words to use in a query, since researchers tend to use make up names for new ideas.

Approach

Users submit entire paper manuscripts as queries
 We use the citation structure of the literature in order to identify related and important papers.

Evaluation and Results

Evaluated with research papers from the Rexa corpus
 Performed manual evaluation to check the approach
 Performed automatic evaluation by stripping bibliographies from real research papers, using them as queries, then evaluating our ranked list versus the paper's real bibliography.
 Our experimental model significantly outperforms a text similarity baseline, and the Katz graph distance measure is the most important additional feature.

		Full			Truncated		
		Mean	Confidence	Interval	Mean	Confidence	Interval
Baseline	Text Similarity	0.0079	0.0055	0.0103	0.0079	0.0055	0.0103
Experimental Models	All Features	0.1016	0.0781	0.1251	0.0940	0.0727	0.1153
	No Text	0.0675	0.0539	0.0811	0.0612	0.0469	0.0754
	No Author	0.0983	0.0747	0.1219	0.0917	0.0701	0.1132
	No Katz	0.0335	0.0256	0.0414	0.0257	0.0194	0.0320
	No Cite Count	0.1005	0.0771	0.1238	0.0931	0.0718	0.1144
	No Date	0.1052	0.0834	0.1269	0.0979	0.0784	0.1174
	No Title	0.1016	0.0781	0.1251	0.0940	0.0727	0.1153

Total paper entries	964,977
Papers with text	105,601
Total citations	1.46 million
Total cited papers	675,372

Statistics from the Rexa corpus

Publication year	Year the paper was published, normalized by subtracting 1950
Text similarity	Similarity of the text of this candidate with the query, as measured by the multinomial diffusion kernel
Co-citation coupling	The fraction of the documents that cite this candidate that also cite documents in the base set
Same author	True if this document is written by the same author that wrote the query
Katz	The Katz graph distance measure
Citation count	Number of citations of this document from all documents in the corpus

Features used in ranking