
Re-ranking Search Results based on Perturbation of Concept-Association Graphs

Gaurav Chandalia and Rohini Srihari
Computer Science, SUNY – University at Buffalo



With thanks to Matthew Beal for discussions on Matrix
Perturbation theory

Objective

- Capture the **semantics** of the document corpus
- Explore the **inherent relationships** between concepts (words) across all the documents
- Effective algorithm that can exploit such a semantic representation to **re-rank search results**

Motivation

- For advanced IR applications like question answering, traditional *bag of words* model is not enough
- Semantic Representation
 - How such information can be leveraged and used as building blocks for advanced IR applications
- Technique should...
 - Not be explicitly dependent on the query
 - Be domain independent

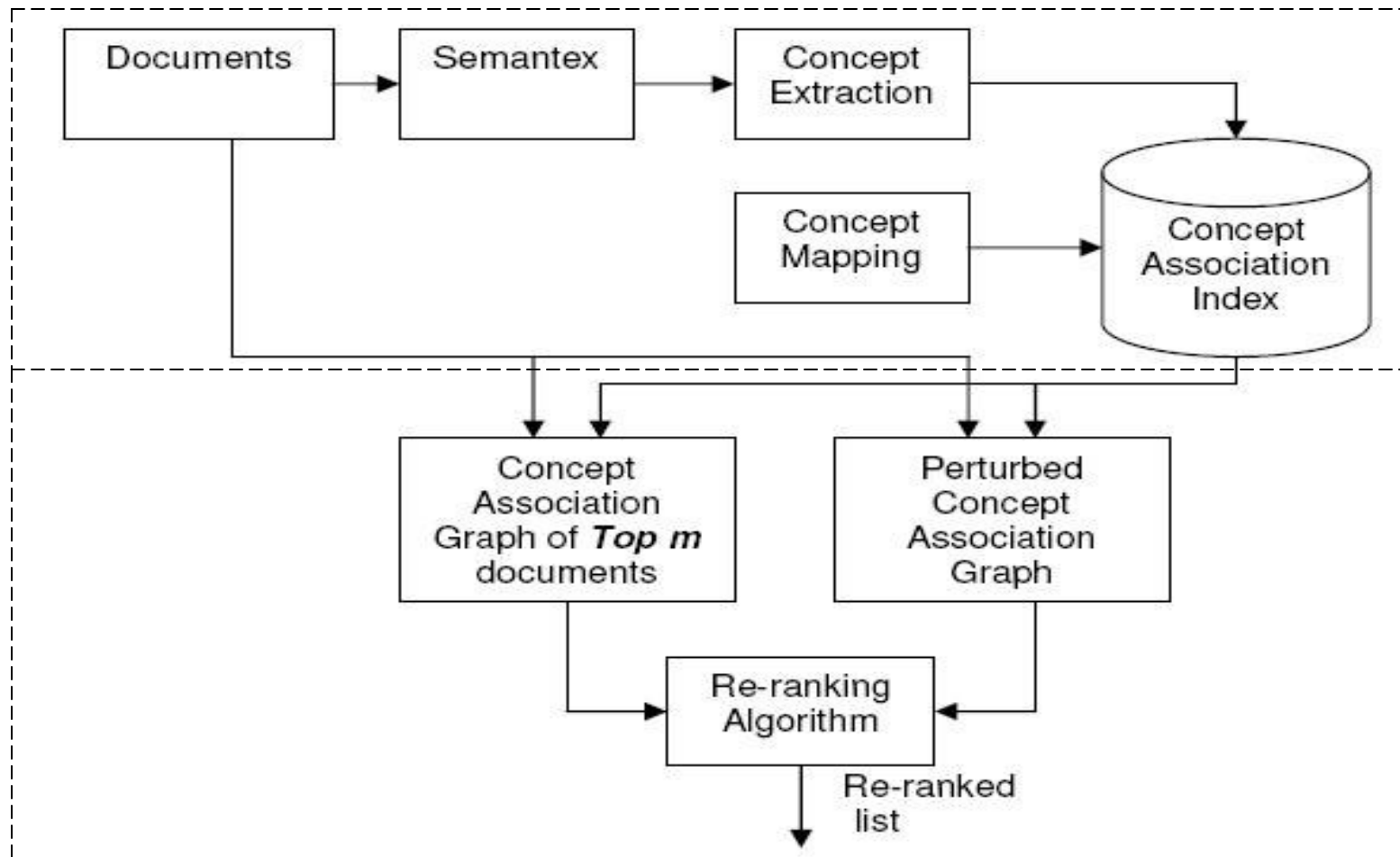
Approach

- Process the document corpus to extract **salient** concepts and associations between the terms (offline)
- For a given query, retrieve the documents based on the Vector Space Model
- Consider the Concept-Association graph of a relevant subset of search results
- Perturb the graph and see if the resulting graph is still relevant to the query
- Re-rank the search results based on the amount of perturbation introduced in the original graph
- Raises two questions...
 - What is the nature of perturbation?
 - How to measure the change in the graph after perturbation?

Approach...

- Nature of Perturbation
 - Perturb the original graph by adding associations between certain concepts
 - Associations are obtained from documents that are to be re-ranked
- Measure the change after Perturbation
 - Perturbed Subspace HITS algorithm
 - Considers projection of eigenvectors on subspace representing the original graph

System Overview



Concept-Association Graph

- Concepts
 - Named Entity objects representing items such as names of *person, organization, etc...*
 - Noun Groups
 - Arguments of General Events
- Associations
 - Capture relationships between the concepts such as *affiliation* associated with a person
 - General Events such as *Subject-Verb-Object* patterns
- Graph Construction
 - Represent the Concept-Association graph as an adjacency matrix of $n \times n$ concepts
 - Association between two concepts (i,j) is indicated by making the $(i, j)^{th}$ entry in the matrix **1** and **0** otherwise

Re-ranking search results

- Perform link analysis on the Concept-Association graph to measure the effects of perturbation
- Algorithm is based on...
 - Hypertext Induced Topic Selection (HITS)
 - Identify *authoritative* web pages
 - Represent the Web as an adjacency matrix and use the iterative power method to compute the principal eigenvectors of the matrix
 - Subspace HITS
 - Identify *authoritative* web pages by projecting each eigenvector representing a hub or an authority on a *subspace*

Perturbed Subspace HITS algorithm

- Compute the top k eigenvectors $V^a = (\mathbf{v}_1^a, \mathbf{v}_2^a, \dots, \mathbf{v}_k^a)$ and the eigenvalues $d_1^a, d_1^a, \dots, d_k^a$ of $T = A^T A$, where A is the adjacency matrix of the original graph
- Compute the **authority scores** of $T = A^T A$ as

$$x_s^a = \sum_{p=1}^k d_p^a ((\mathbf{v}_s^a)^T \mathbf{v}_p^a)^2 \quad \text{where } s = 1 \text{ to } k$$

- For each document doc that is to be re-ranked
 - Initialize $B = A$. **Perturb** the graph B using the concepts and associations from doc
 - Compute the top k eigenvectors $V^b = (\mathbf{v}_1^b, \mathbf{v}_2^b, \dots, \mathbf{v}_k^b)$ and the eigenvalues $d_1^b, d_1^b, \dots, d_k^b$ of $S = B^T B$
 - Compute the **change in authority scores** for document doc with respect to the original graph in the following way:

$$\mathbf{f}_s = \sum_{p=1}^k d_p^b ((\mathbf{v}_s^b)^T \mathbf{v}_p^a)^2 - x_s^a \quad \text{where } s = 1 \text{ to } k$$

Perturbed Subspace HITS algorithm...

- Re-rank the documents using *Borda Count*
 - Vectors representing the change in authority scores act as *voters* and documents act as *candidates*
 - Each *voter* ranks all the *candidates*
 - Ranks from all the *voters* are then combined to give a single ranked list of documents

Dataset

- High Accuracy Retrieval of Documents (*HARD*) track - 2003 *Text REtrieval Conference (TREC)*
- Results were evaluated using the standard TREC Evaluation Code

	NYT	APW	XIE	CR	FR	Total
No. of docs.	137,806	77,876	104,698	16,609	35,230	372,219
Size	750MB	245MB	310MB	147MB	330MB	1.7GB

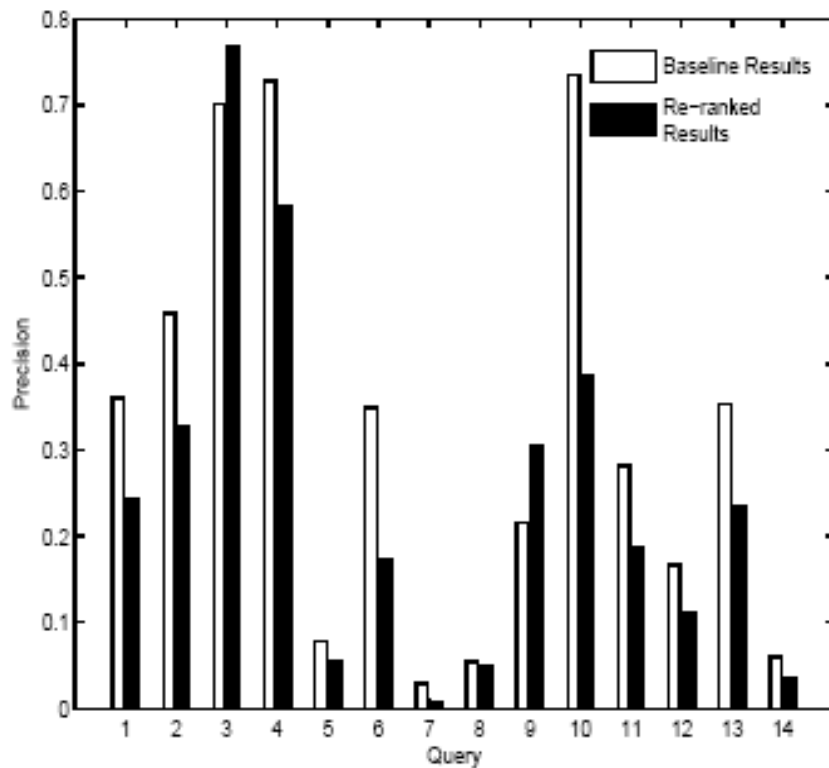
Dataset...

Query No.	Query	Relevant Docs. in the collection	Docs. Retrieved in baseline run
33	Animal Protection	401	211
48	Y2K Crisis	562	301
51	Hate Crimes Prevention	168	140
65	Mad Cow Disease	145	126
69	Environmental Protection	513	101
70	Red Cross activities	111	90
77	Insect-Borne illnesses	194	25
84	Recent Earthquakes	86	43
99	Globalization and Democracy	399	170
102	Microsoft monopoly	285	249
116	Genetic Modification technology	200	116
146	NATO/UN Tension over Balkans crisis	305	131
147	Regional Economic integration	327	181
187	National Leadership Transactions	194	51

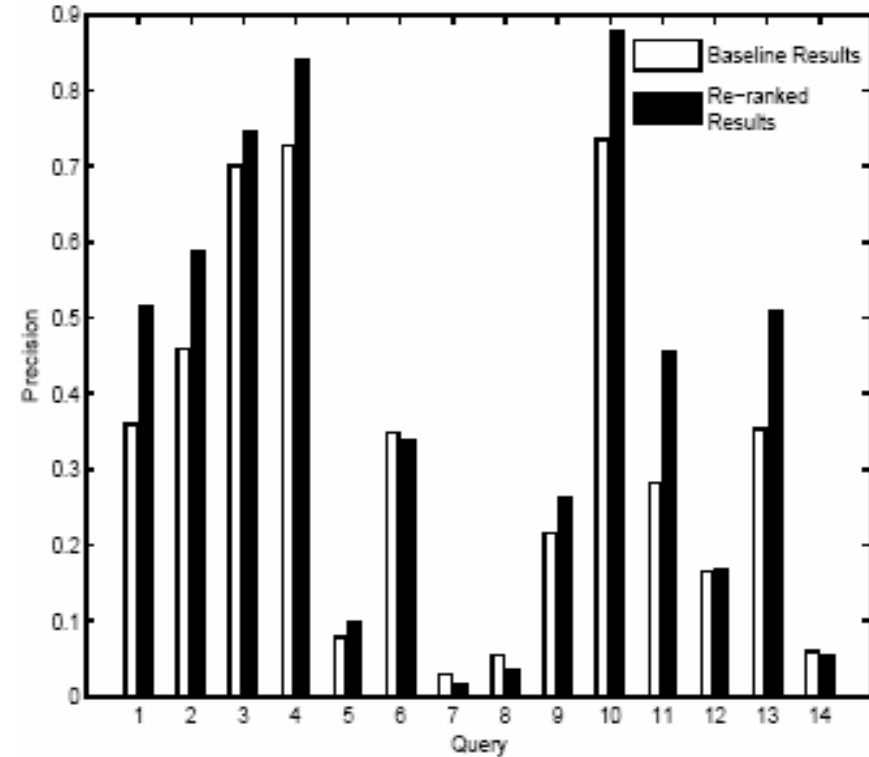
Results

- Query vs. Exact Precision graph

HITS



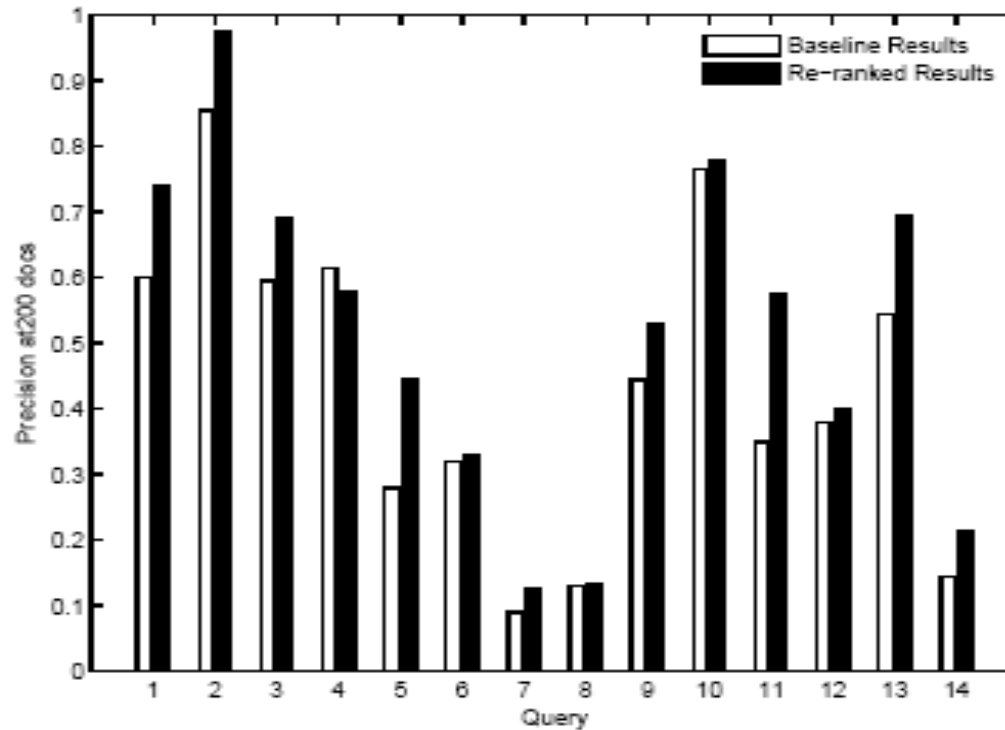
Perturbed Subspace HITS



Results...

- Query vs. Exact Precision at 200 documents retrieved graph

Perturbed Subspace HITS



Conclusions & Future Work

- Conclusions
 - Our approach is not explicitly dependent on the query
 - Exploits the information contained in the associations (semantic links) between concepts (words)
 - Re-ranking technique is based on blind feedback
- Future Work
 - Evaluation on a larger set of queries
 - Formalize the Perturbed Subspace HITS re-ranking algorithm
 - Improve the extraction of concepts and associations

References

- J. Aslam and M. Montague. **Models for metasearch.** *Proc. SIGIR 2001*, September 2001.
- J. Kleinberg. **Authoritative sources in a hyperlinked environment.** In *Proc. 9th Ann. ACM SIAM Symp. Discrete Algorithms*, pages 668–677. ACM, 1998.
- A. Y. Ng, A. X. Zheng, and M. I. Jordan. **Link analysis, eigenvectors and stability.** In *Proc. 17th International Joint Conference on Artificial Intelligence*, 2001.
- A. Y. Ng, A. X. Zheng, and M. I. Jordan. **Stable algorithms for link analysis.** In *Proc. 24th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.
- G. W. Stewart and J. Guang Sun. **Matrix Perturbation Theory.** *Academic Press, Inc.*, London, UK, 1990.
- R. K. Srihari, W. Li, C. Niu, and T. Cornell. **Semantex: A customizable intermediate level information extraction engine.** To appear in the *Journal of Natural Language Engineering*, 12(4), 2006.