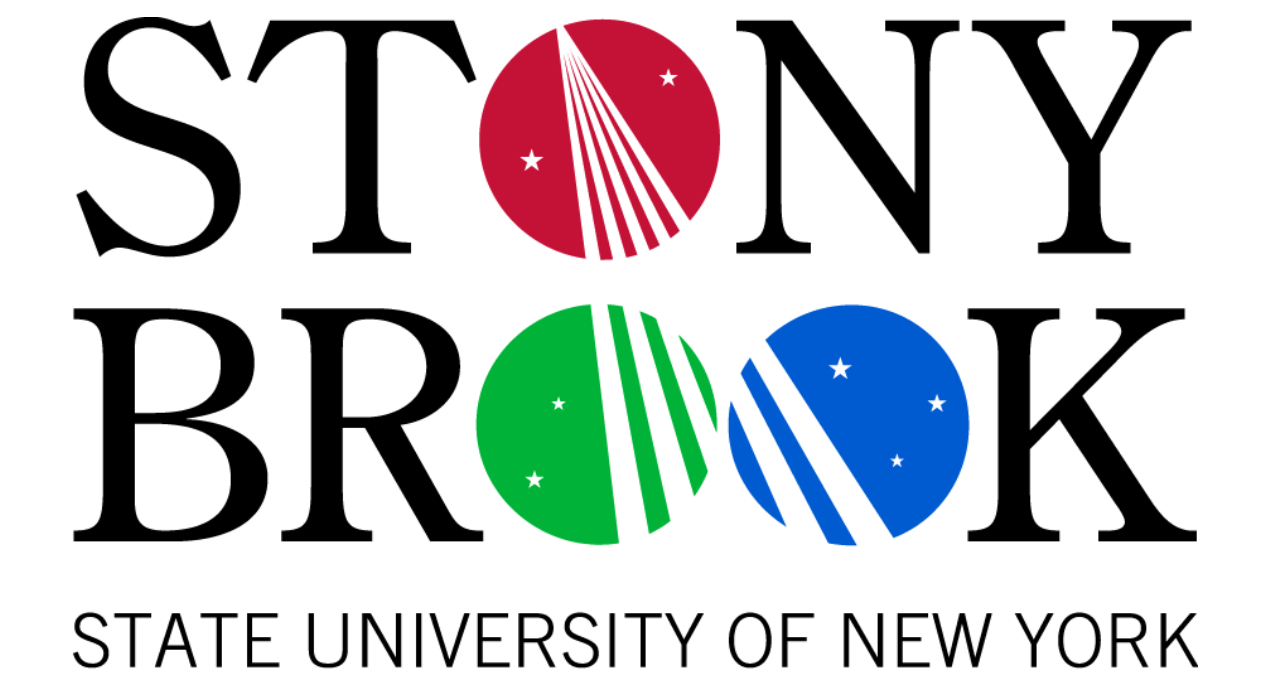


# Concordance-Based Entity-Oriented Search

Mikhail Bautin and Steven Skiena

Department Of Computer Science, Stony Brook University, Stony Brook, NY 11794-4400



## Introduction

We consider the problem of finding the relevant named entities in response to a search query over a given text corpus. We believe it would be quite valuable for users to get lists of the “most relevant” entities to their query in addition to the most relevant documents. Our entity search engine is part of the Lydia news analysis system available at <http://www.textmap.com>.

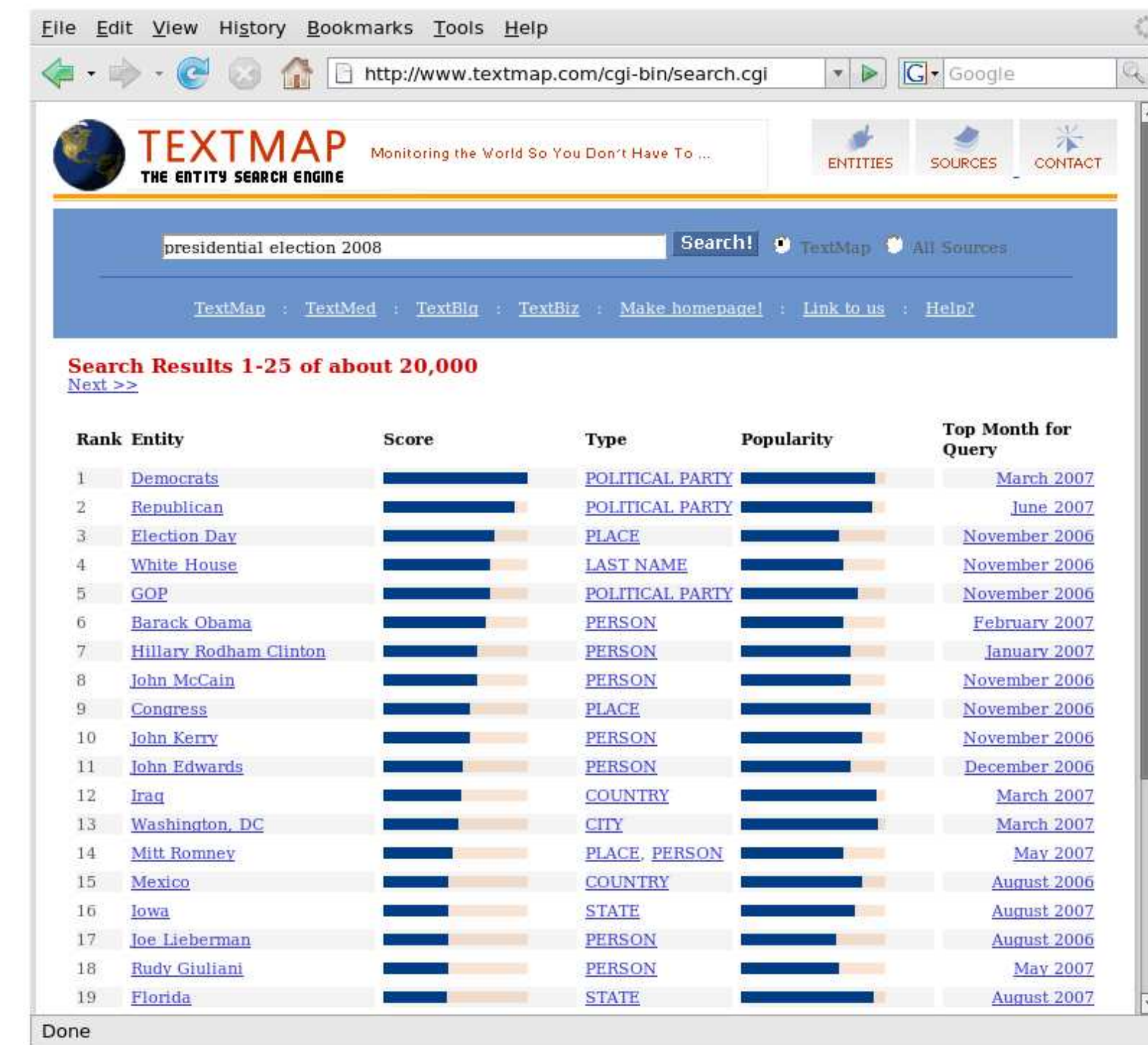
	TENNIS	NEW YORK YANKEES	GOOGLE
1	Tennis	New York Yankees	Google
2	Roger Federer	Joe Torre	Yahoo
3	Andre Agassi	Alex Rodriguez	Eric Schmidt
4	U.S. Open	Derek Jeter	Mountain View
5	Andy Roddick	Boston Red Sox	Microsoft

	POLITICAL CORRUPTION	POLARIZING FIGURE	BRITISH PRIME MINISTER
1	Jack Abramoff	Hillary Rodham Clinton	Tony Blair
2	George Ryan	Katherine Harris	Winston Churchill
3	Tom DeLay	David Geffen	Margaret Thatcher
4	Pete Domenici	Donald H. Rumsfeld	British
5	King Gyanendra	Dick Cheney	Gordon Brown

*Example of entity queries and results.*

Possible applications include:

- **Navigational Search.** The related entity results may satisfy the user’s information need, as in question answering systems, or provide meaningful navigational alternatives to user’s document-oriented query.
- **Encyclopedia Search.** Encyclopedias are inherently entity-oriented, so our techniques can be readily applied to them. We are confident our system could improve Wikipedia search.
- **Product Search.** Aggregating all mentions of specific products in reviews, blogs and webpages can give a higher recall than existing product search engines do.
- **Search Recommendations.** Entity search results are useful as “you may also try” hints. Our experiments show that for certain types of entity queries the next query by the same user appears within the top entity search results of the previous query in 4-8% cases.



## Entities in Web Search Logs

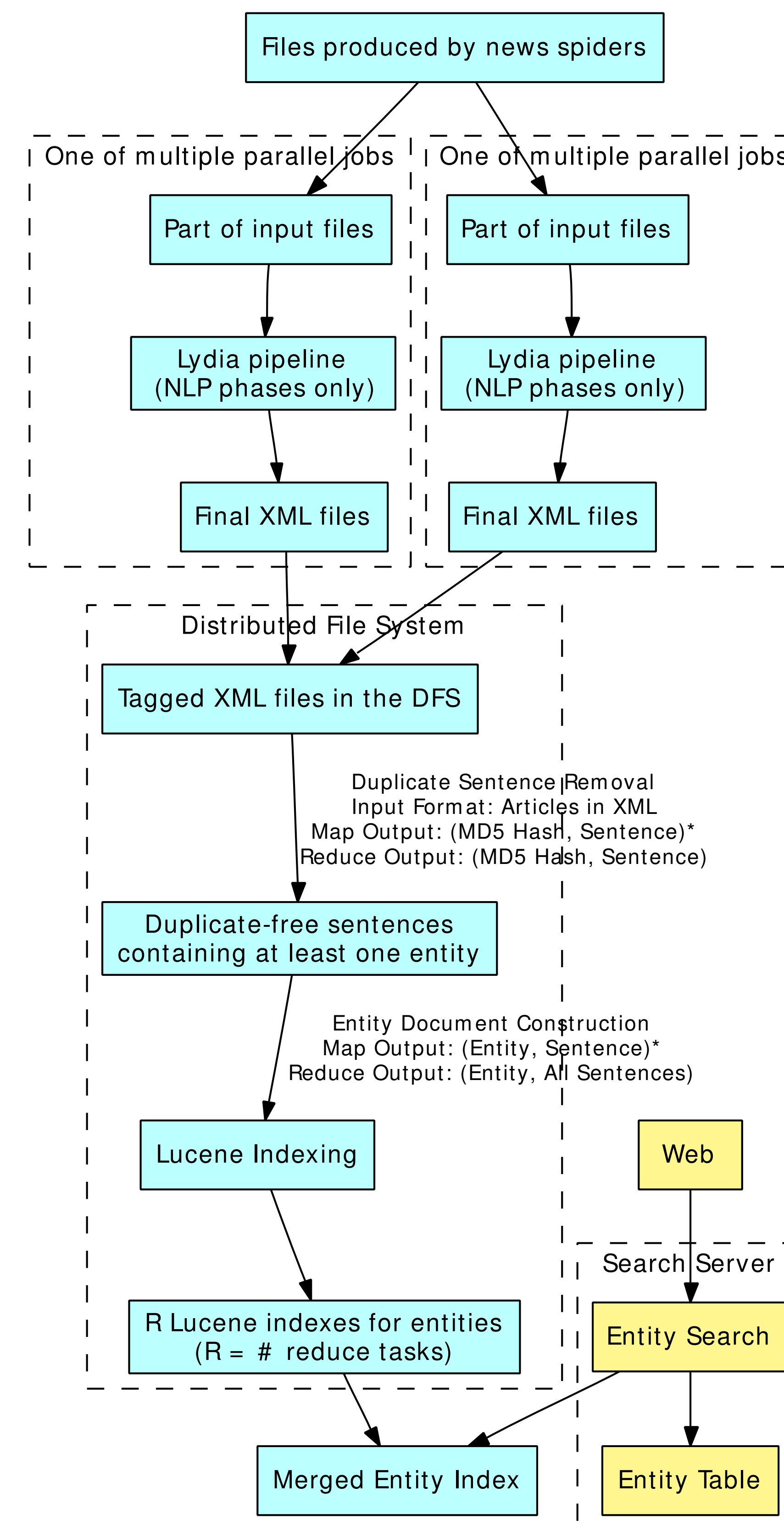
	All Queries		
Matches	No aliases	Aliases	Metaphone
perfect	17.91%	26.50%	38.82%
partial	55.14%	53.59%	48.41%
total	73.05%	80.09%	87.23%

	Unique Queries		
Matches	No aliases	Aliases	Metaphone
perfect	2.07%	5.33%	18.57%
partial	68.85%	69.72%	65.23%
total	70.92%	75.05%	83.80%

*Match frequencies for all 36,389,577 queries and 10,154,743 unique queries (after duplicate removal) compared against Lydia entity list.*

Seeking motivation for entity-oriented search we asked the question: how often do web search queries recognizably target entities? We used web search query dataset released by AOL in August 2006. When analyzing it, we did not reveal user identity or manually examine individual low-frequency queries, thus maintaining and respecting user privacy. The above table gives comparison results of these queries to entity lists obtained from our Lydia news analysis system and Wikipedia.

## Implementation



Indexing phase steps:

1. Processing news articles with the Lydia pipeline.
2. Removing duplicate sentences.
3. Collecting all sentences containing each entity into a “concordance” document for each (entity, month) pair.
4. Indexing concordance documents with Lucene.

During the search entity scores are calculated from the (entity, month) pair scores returned by Lucene.

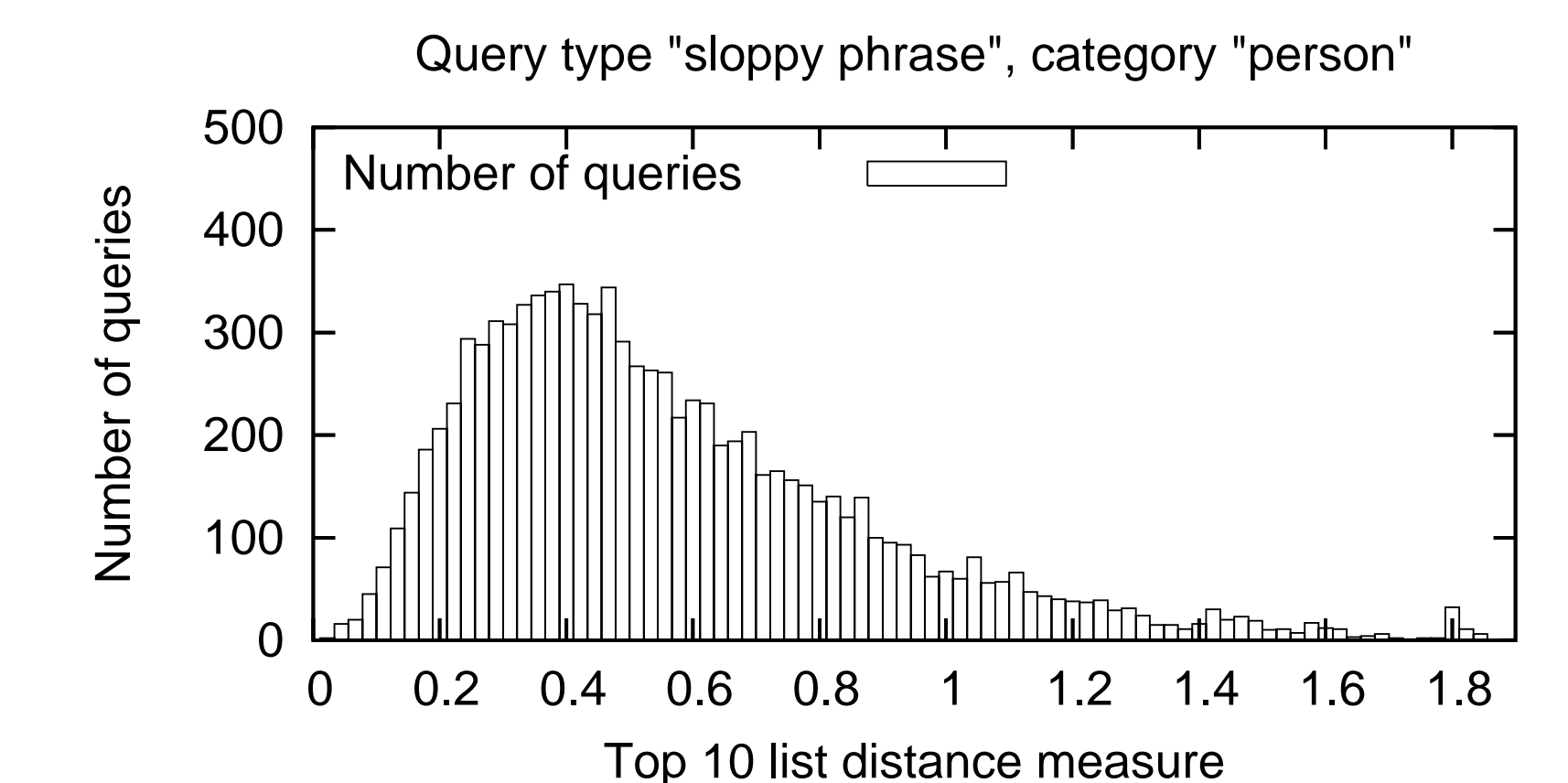
## Evaluation

- **Comparison with juxtaposition score.** Juxtaposition score is an existing entity relatedness measure in the Lydia system:

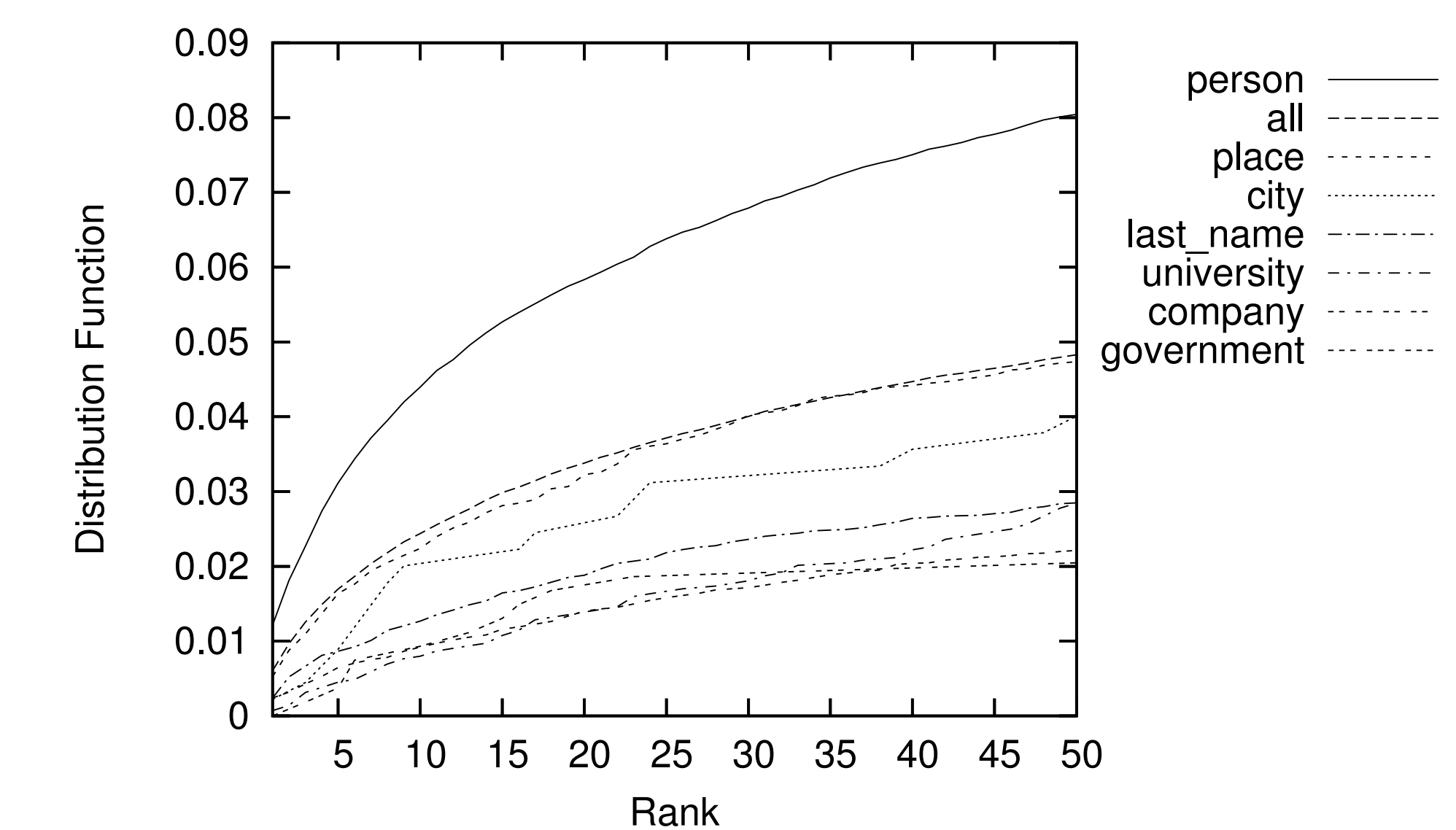
**Juxtapositions for Hillary Rodham Clinton:** [What is this?](#) [More...](#)



To compare this with the results of our entity search engine for an entity given as query, we use Kendall’s distance measure for top-*k* lists. The best similarity is achieved for the category of people.



- **Search recommendations: predicting user’s next query.** For a given user web query that is an entity (using AOL data), we measure how deep his or her next query appears in our list of entity results for the previous query.



Again, we achieve the best performance for the category of people: in 4% cases the next query is in our top-10 list and in 8% cases it is in the top-50 list.