

EECS 395/495 Lecture 6: Document Ranking and Link Analysis

Doug Downey

How Does Web Search Work?

- Inverted Indices
- Web Crawlers
- Ranking Algorithms
 - Basics of Machine Learning and Text Classification
 - **Document Ranking**
 - **Link Analysis**

Document Ranking

- Static Features
 - Which Web pages are good?
- Dynamic Features
 - Which pages are most relevant to the query?

Static Ranking (1)

- Known Reputable Sites
 - Wikipedia, .edu domains, etc.
- Spam detection
- Link Analysis
 - More later
- (Real) Popularity
- Page features
 - Length of page, length of URL
- Anchor Text features
 - How much text in links to page, how varied,...

Static Ranking (2)

Table 2: Results for individual feature sets.

<i>Feature Set</i>	<i>Accuracy (%)</i>
PageRank	56.70
Popularity	60.82
Anchor	59.09
Page	63.93
Domain	59.03
All Features	67.43

- Beyond PageRank: Machine Learning for Static Ranking
Richardson et al., 2005

Dynamic Ranking

- Vector-space model
- Term frequency (tf) vectors:

	d1	d2	d3	d4	d5	d6
airplane	2	21	0	0	0	10
does	3	20	13	10	2	12
elephant	2	3	0	2	0	0
found	12	4	2	3	12	1
house	1	1	0	0	1	0

- The “bag of words” representation
- Idea: $\text{score}(q, d) = \mathbf{q} \cdot \mathbf{d}$
(i.e. dot product of frequency vectors)

Normalization

- Problem with naïve vector space model
 - Unimportant terms dominate
- Solution: Normalize frequencies by
 - Overall frequency in the document collection?
 - Number of distinct docs mentioning term?

Word	cf	df
try	10422	8760
insurance	10440	3997

TF-IDF Weighting

- tf: term frequency
- idf: inverse document frequency
 - $\log (N / df)$
N = total # of documents

$$\text{Score}(q, d) = \sum_{t \in q} \text{tf-idf}_{t,d}$$

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008

On the Web: Other Weighting Factors

- Title
- Anchor Text
- Font/Color
- Position on page
- **Implicit Feedback**

Putting it all together

- Google's ~400 lines of code (?)
- Microsoft Live Search
 - Static & dynamic features computed for several query/document sets
 - Previously: Neural network trained on human-annotated result ratings

Table 1: Basic Results

<i>Technique</i>	<i>Accuracy (%)</i>
None (Baseline)	50.00
PageRank	56.70
fRank	67.43

How Does Web Search Work?

- Inverted Indices
- Web Crawlers
- Ranking Algorithms
 - Basics of Machine Learning and Text Classification
 - Document Ranking
 - **Link Analysis**