
Searching the Workplace Web

Ronald Fagin Ravi Kumar Kevin S.McCurley
Jasmine Novak D. Sivakumar
John A.Tomlin David P.Williamson

Presentation by Na Dai

Motivation

- Influence of social forces on internet vs. intranet
 - Reflections
 - Development guidance
 - Difference measure of success
- Few research on intranet search
 - Corporations don't want to expose their intranet
 - Limited privileges to access intranet

Share a great deal, but we need to consider the unique characteristics of intranet search!

Core Ideas

- Focus on: Intranet web search ranking problem
 - Characteristics: Robustness and flexibility
 - Decouple the ranking process
 - Selection of ranking heuristics
 - Synthesis of ranking methods
-

Comparison on Nature: Internet vs. Intranet

- Axiom 1. Intranet documents are often created for simple dissemination of information, rather than to attract and hold the attention of any specific group of users.
 - Axiom2. A large fraction of queries tend to have a small set of correct answers (often unique), and the unique answer pages do not usually have any special characteristics.
 - Axiom3. Intranets are essentially spam-free.
 - Axiom4. Large portions of intranets are not search-engine-friendly.
-

Difference on graph structures: Internet vs. Intranet (1)

■ Generalization:

- Heterogeneous
- Diverse: 7000 hosts
- All kinds of commercial web servers

■ Specialty:

- Lotus Domino servers

- Estimated: 50M URLs
 - Crawled: 20M URLs
 - After delete Duplicate Links: 4.6M URLs
 - 3.4M pages (containing anchortext)
-

Difference on graph structures: Internet vs. Intranet (2)

- Indegree and outdegree distributions
- Connectivity properties
 - Internet:
 - SCC: 30%
 - IN: 25%
 - OUT: 25%
 - IBM Intranet
 - SCC: 10%
 - OUT: large

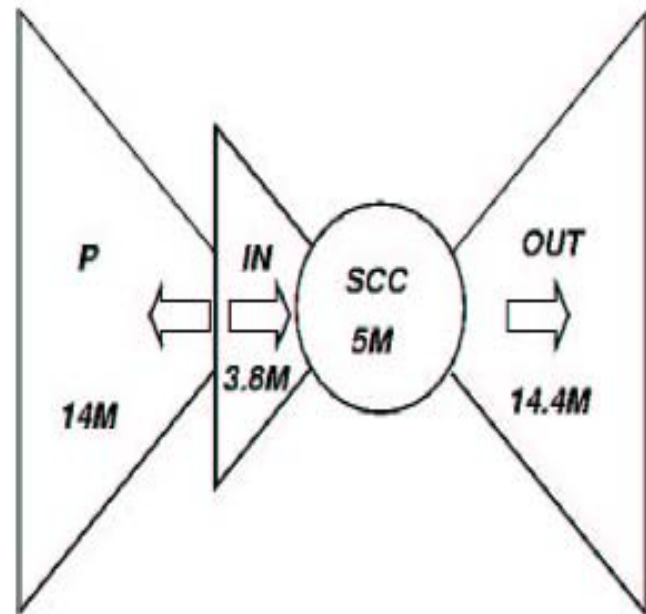


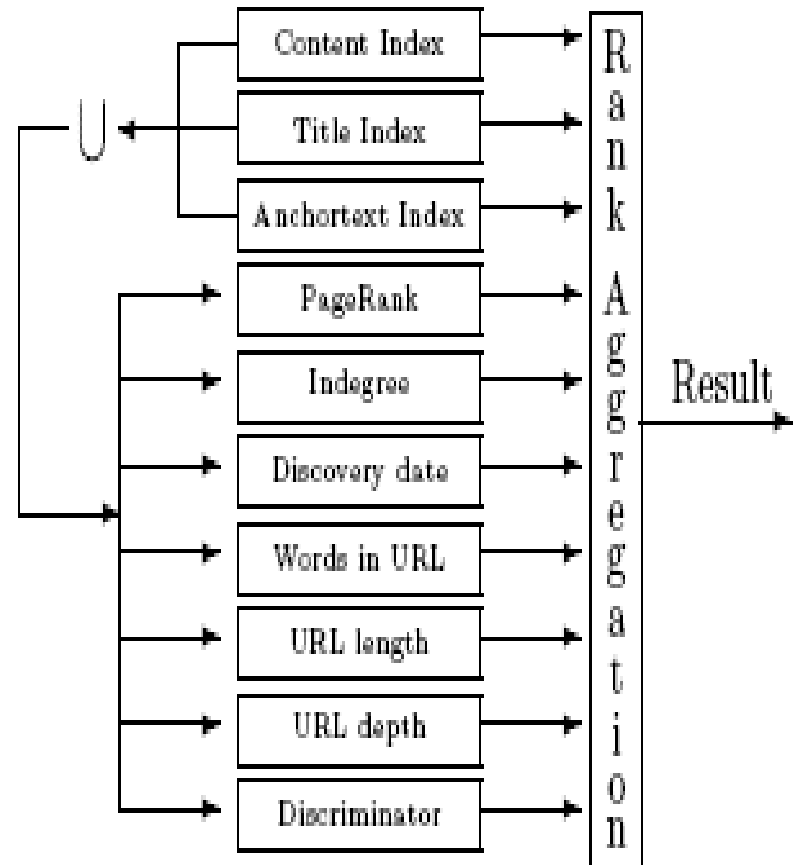
Figure 1: Macro-level connectivity of IBM intranet

System Architecture

- Crawler
 - URLs; metadata; canonicalization
 - Global ranking component
 - Query-independent ranking tables
 - Duplicate elimination component
 - use shingle, group, select one favorite for indexing
 - Inverted index engine
 - Primary copies of documents
 - Multiple separate indices : content, title, anchortext
 - Separate dictionaries
 - Query runtime system
 - Result markup and presentation system
-

Rank Aggregation

- Aim: minimize the total number of inversion
- Kendall tau distance
 - $K(\sigma, \pi) = \# \text{ unordered pair}\{k, l\}$
 - $K_j(\sigma, \tau_j) = |S_j(\sigma, \tau_j)| / \binom{|U_j|}{2}$
 - minimizes $\sum_j K_j(\sigma, \tau_j)$
- MC4 (Markov chains)



Experiment

- Query sets:
 - Q1: 131 top popular queries (broad topic, single-word, directed towards “hubs”)
 - Q2: 82 queries with median frequency (specific, longer, “typical” queries)
- What is the correct answers for queries?
 - Existing search engine + good old browsing
 - Locate seed answer + refined by browsing
 - Ambiguous? Multiple answers.
- Evaluation Criteria
 - Recall at position p
 - $I_{\mu}(\alpha) = (\mu(C^+) - \mu(C^-)) / \mu(C^-)$
 - Similarity : Kmin

Experiment Results

α	$IR_1(\alpha)$	$IR_3(\alpha)$	$IR_5(\alpha)$	$IR_{10}(\alpha)$	$IR_{20}(\alpha)$
Ti	29.2	13.6	5.6	6.2	5.6
An	24.0	47.1	58.3	74.4	87.5
Co	3.3	-6.0	-7.0	-4.4	-2.7
Le	3.3	4.2	1.8	0	0
De	-9.7	-4.0	-3.5	-2.9	-4.0
Wo	3.3	0	-1.8	0	1.4
Di	0	-2.0	-1.8	0	0
PR	0	13.6	11.8	7.9	2.7
In	0	-2.0	-1.8	1.5	0
Da	0	4.2	5.6	4.6	0

Table 1: Influences of various ranking heuristics on the recall at various positions on the query set Q_1

α	$IR_1(\alpha)$	$IR_3(\alpha)$	$IR_5(\alpha)$	$IR_{10}(\alpha)$	$IR_{20}(\alpha)$
Ti	6.7	8.7	3.4	3.0	0
An	23.1	31.6	30.4	21.4	15.2
Co	-6.2	-4.0	3.4	0	5.6
Le	6.7	-4.0	0	0	-5.3
De	-18.8	-8.0	-10	-8.8	-7.9
Wo	6.7	-4.0	0	0	0
Di	-6.2	-4.0	0	0	0
PR	6.7	4.2	11.1	6.2	2.7
In	-6.2	-4.0	0	0	0
Da	14.3	4.2	3.4	0	2.7

Table 2: Influences of various ranking heuristics on the recall at various positions on the query set Q_2

- Information in anchor text, doc titles, keyword descriptors and meta-data is valuable for intranet search
- Building separate indices is effective
- Information from indices is query-independent.

Experiment Results

α	$I_{R1}(\alpha)$	$I_{R3}(\alpha)$	$I_{R5}(\alpha)$	$I_{R10}(\alpha)$	$I_{R20}(\alpha)$
Ti	29.2	13.6	5.6	6.2	5.6
An	24.0	47.1	58.3	74.4	87.5
Co	3.3	-6.0	-7.0	-4.4	-2.7
Le	3.3	4.2	1.8	0	0
De	-9.7	-4.0	-3.5	-2.9	-4.0
Wo	3.3	0	-1.8	0	1.4
Di	0	-2.0	-1.8	0	0
PR	0	13.6	11.8	7.9	2.7
In	0	-2.0	-1.8	1.5	0
Da	0	4.2	5.6	4.6	0

Table 1: Influences of various ranking heuristics on the recall at various positions on the query set Q_1

α	$I_{R1}(\alpha)$	$I_{R3}(\alpha)$	$I_{R5}(\alpha)$	$I_{R10}(\alpha)$	$I_{R20}(\alpha)$
Ti	6.7	8.7	3.4	3.0	0
An	23.1	31.6	30.4	21.4	15.2
Co	-6.2	-4.0	3.4	0	5.6
Le	6.7	-4.0	0	0	-5.3
De	-18.8	-8.0	-10	-8.8	-7.9
Wo	6.7	-4.0	0	0	0
Di	-6.2	-4.0	0	0	0
PR	6.7	4.2	11.1	6.2	2.7
In	-6.2	-4.0	0	0	0
Da	14.3	4.2	3.4	0	2.7

Table 2: Influences of various ranking heuristics on the recall at various positions on the query set Q_2

- For different type of queries, different heuristics have different performances. (classifier? Learn from logs)
- Many auxiliary heuristics are quite useful

Experiment Results

α	AG	Ti	An	Co	Le	De	Wo	Di	PR	In	Da
AG	0	26	34	26	32	42	47	42	36	34	13
Ti	26	0	27	21	34	31	26	29	27	31	08
An	34	27	0	32	45	46	40	43	41	33	13
Co	26	21	32	0	21	31	36	35	33	37	12
Le	32	34	45	21	0	33	54	43	51	49	17
De	42	31	46	31	33	0	53	52	51	52	20
Wo	47	26	40	36	54	53	0	48	48	49	16
Di	42	29	43	35	43	52	48	0	47	46	15
PR	36	27	41	33	51	51	48	47	0	31	12
In	34	31	33	37	49	52	49	46	31	0	13
Da	13	08	13	12	17	20	16	15	12	13	0

Table 3: Distances between the heuristics and to their aggregation, query set $Q_1 \cup Q_2$, normalized to be between 0 and 100

- 7 auxiliary ranking heuristics are much more closely aligned with the ranking based on the title index
- Apart from Da, auxiliary ranking heuristics are dissimilar

Conclusion

- Difference between Intranet and Internet searches
 - Queries
 - Notion of “good answers”
 - Social process that create Intranet vs. the one that creates Internet
 - Rank aggregation:
 - flexible and modular → follow “plug and play” mode
 - Combine multiple ranking heuristics
-

Thank you!
