Internet search and metasearch

Ravi Kumar
IBM Almaden Research Center
ravi@almaden.ibm.com
http://www.almaden.ibm.com/cs/people/ravi
Overview

I: Internet search
II: More search
III: Metasearch and rank aggregation
I: Internet search
Roadmap

• Classic information retrieval
• IR and the web
• Search engines
• Ranking
• PageRank
• HITS/Clever
• Finding related pages
Goal

- How to search/query/mine the web?

The web

Billions of pages
Tens of billions of links
Constantly changing
Wealth of information
Classic information retrieval (IR)

Input: Set of documents

Goal: Given a query, retrieve documents that are 'relevant' to the query

Method
  – Preprocess the documents
  – Search at query-time
IR models

• Logical model
  – String matches, AND/OR/NOT

• Vector space model
  – Documents/query are vector of terms
  – i-th entry = function of i-th term occurrence in the document
  – Similarity measure between document and query
  – Order documents based on similarity to query

• Probabilistic models, …
What is different about the web?

- **Volume**
  - Few billion pages, few tens of billion links
- **Change**
  - 23%/day, dynamic pages
- **Decay**
  - Short half-life
- **Heterogeneity**
  - HTML pages, pdf/ps/word documents, images, language
What is different… (contd.)

• Duplication
  – Exact, near-duplication, semantic duplication

• Variable quality
  – NASA vs. XYZ’s description of space exploration

• Spam
  – Good source vs. malicious source (‘miserable failure’)

• Links
  – Malicious links, dead links, redirections, dead-ends
What is different... (contd.)

- **User behavior**
  - Poor queries, short, imprecise, badly formed, low effort
  - Focus on top few results

- **Expectation**
  - Instant response

- **Performance evaluation**
  - Is there an absolute truth?
The bright side

• **Structure**
  – Links/HTML
  – Redundancy
  – Enthusiastic and free-spirited content-creators and link-creators

• **Search experience**
  – Personalization
  – Interaction
Internet search engines

- General-purpose search engines
  - Google, MSN, Teoma, Alltheweb, …
- Directories/taxonomies (hand-built)
  - Yahoo!, OpenDirectory
- Special-purpose search engines
  - Travelocity, Orbitz, Addall
- Search by example
  - Related pages in Google, …
- Metasearch
  - Metacrawler
Components of a search engine

- **Crawler**
  - Gathers pages from the web
- **Indexer**
  - Stores the pages in a structured manner (index)
- **Query interface**
  - Serves queries by consulting the index
Algorithmic issues: Crawler

- Load balancing
- Prioritization
- Coverage
- Spam/porn filtering
- Avoid spider traps
- ...
Algorithmic issues: Indexer

- Storage representation
- Duplication elimination
- Template detection
- Query-independent ranking
- Classification/clustering
- …
Algorithmic issues: Query interface

- Query-dependent ranking
- Duplicate elimination
- Query refinement options
- Categorization/clustering
- Related links
- Ads! 😊 (‘Work at Google’)
Ranking

• Input: Web pages
• Goal: Given query, output answers in order of `relevance’ to the query
• Paradigms
  – Query-independent
    • Eg, last modified date, number of citations
  – Query-dependent
    • Eg, cosine similarity
  – Combined
Approaches to ranking

- **Text-based ranking**
  - Classical IR-style
  - Query-dependent

- **Link-based ranking**
  - Query-independent
    - PageRank [Brin Page 1998]
  - Query-dependent
    - HITS [Kleinberg 1998]
Link-analysis in hyperlinked corpus

Citation analysis of scholarly publications
(Bibliometrics)

- Impact factor
- Influence weights
Impact factor

- [Garfield 1972]
- Rank by indegree (per article over past two years)

- Limitations: Not all citations are equally born
- Important journals are ones cited by other important journals
Influence weights

- [Pinksi Narin 1976]
- Citation strength $A_{ij}$ from journal $i$ to journal $j = \text{fraction of citations in } i \text{ that go to } j$

- Influence weight of journal $j = \text{sum of influences of all journals citing } j \text{ scaled by citation strengths}$

$$w_j = \sum_i A_{ij} w_i$$

$$w = A^T w, \text{ the eigenvector of } A \text{ associated with eigenvalue } 1$$
Eigenvectors and link structures

- **Random walks**
  - Begin at a random node
  - At each step, move to a neighbor with indicated probability \( (M_{ij}) \)
- **Theorem**: If the graph is strongly connected and not periodic, then there is a unique stationary distribution \( p \)
- \( M^T p = p \), \( p \) is eigenvector associated with eigenvalue 1
Hypertext IR principles

• Relevant linkage principle
  – p links to q ⇒ q is relevant to p

• Topical unity principle [Kessler 1963, Small 1973]
  – q₁ and q₂ are co-cited in p
    ⇒ q₁ and q₂ are related to each other

• Lexical affinity principle [Maarek et al. 1991]
  – Closer the links to q₁ and q₂ are, stronger the relation between them
Web as a directed graph

- Nodes = (static) web pages
- Directed edges = edge from p to q if page p has a hyperlink to page q
Query-independent ranking

- Page \( P \) pointing to page \( Q \) = endorsement of page \( Q \) by page \( P \)
- Quality of \( P \) = number of endorsements it receives = indegree of \( P \) in the web graph
- Quality of \( P \) depends on
  - Indegree of \( P \)
  - Quality of pages pointing to \( P \)
- A recursive definition!
PageRank [Brin Page 1998]

Random walk interpretation

- Walk starts at a uniformly chosen web page
- At each step, if currently at page $P$
  - $W/p \alpha$, go to a uniformly chosen web page
  - $W/p \ 1 - \alpha$, go to a uniformly chosen outneighbor of $P$
- $\text{PageRank}(P) = \text{fraction of steps random walk spends at } P \text{ in the limit}$
Mathematically speaking...

- $A =$ adjacency matrix of web graph
- $PR(u) = \alpha/n + (1 - \alpha) \sum_{v \mid (u, v) \in A} \frac{PR(v)}{\text{outdegree}(u)}$
- $M = \alpha U + (1 - \alpha) A$
- PageRank = stationary probability for this Markov chain

$\alpha = 0.15$
Google and PageRank

- Google is based on PageRank
- Query-independent phase
  - Ranks all pages according to PageRank
- Query-dependent phase
  - Return pages containing the query in order of PageRank
  - Further heuristics (title, anchortext, last update, …)
Variants: Topic-sensitive PageRank

[Haveliwala 2003]

- Captures notion of importance wrt given topic
- Instead of jump to a random page, jump to a page w/p proportional to its relevance to the topic
- $W/p \propto p_v$, jump to $v$, where $p_v =$ relevance of $v$ to the topic
- Can precompute small set of relevant pages and set $p_v$ to be uniform among these pages
Query-dependent ranking

- Page P pointing to page Q = P endorses Q
- But, two popular pages may not cite each other

Two-layer model: Hubs and authorities
- **Hubs**: Pages with pointers to lots of resources for a topic
- **Authorities**: Representative sources for a topic
- Identify the best hubs and authorities for a given topic
HITS [Kleinberg 1998]

- A page is
  - An authority if lots of pages point to it
  - A good authority if lots of pages that are good hubs point to it

- A page is
  - A hub if it points to lots of pages
  - A good hub if it points to lots of pages that are good authorities

A mutually reinforcing and recursive definition!
Mathematically speaking...

- Each page $P$ has $h[P] = a[P] = 1$ initially
- Compute hub scores using authority scores
  \[ h[P] = \sum_{Q ightarrow P} a[Q] \]
- Compute authority scores using hub scores
  \[ a[P] = \sum_{Q ightarrow P} h[Q] \]
- Renormalize scores and repeat
- Output top few hubs and authorities
Mathematically… (contd.)

- \( a_{i+1}(Q) = \sum_{P \rightarrow Q} h_i(P) \), \( h_{i+1}(Q) = \sum_{Q \rightarrow P} a_i(P) \)
- \( a_{i+1} = A^T h_i \), \( h_{i+1} = A a_i \)
- \( a_{i+1} = (A^T A) a_i \), \( h_{i+1} = (A A^T) h_i \)

- Iteration converges to \( a^* \), \( h^* \)

- \( a^* \), \( h^* \) are eigenvectors of \( A A^T \), \( A^T A \)
- \( a^* \), \( h^* \) are left and right singular vectors of \( A \)
Clever [Chakrabarti Dom Gibson Kumar Raghavan Rajagopalan Tomkins 1999]

- Edges in the graph have weights
- Weight is a function of
  - Anchor text vs. query
  - +/- prefixes in the query
  - Source/destination of hyperlink
  - Stop sites
  - Useful sites
HITS/Clever: Implementation

- Apply keyword search to generate initial set of 200 pages
- Expand initial set into root set by following links
- Compute weights for edges
- Perform iterations
- Output top hubs and authorities

- Teoma is based on HITS

Root set

Initial set
User study

- Compare Clever to Altavista/Yahoo (1999)
- 26 query topics
- 10 pages from each source
- Blind test: 37 users, 1369 judgements
Some heuristics

- Two pages from same site contribute less to score
- **Hub functions:** Compute authority scores using per-link hub scores and recompute hub scores using per-page authority scores, but spread weight among neighboring links
- **Covering functions:** Output best set of hubs with less overlap among them
- **Limiting influence:** Weight edges to limit influence
- **Averaging:** Hub score = average of authority scores
  authority score = sum of hub scores that are > average
- **Pagelets**
**Variants of HITS: SALSA**

[Lempel Moran 2000]

**Given a set of pages**
- Out-step (O): Go to a uniform out-link
- In-step (I): Go to a uniform in-link

- **Authority scores = fixed point of O-I chain**
- **Hub scores = fixed point of I-O chain**
- **If \( v \) is in component \( V_v \) with \( E_v \) links**

\[
\alpha(v) = \frac{|V_v|}{|V|} \cdot \text{indegree}(v)/E_v
\]
PageRank vs. HITS/Clever

PageRank
- Query-independent
- Offline computation
- Large graph
- Additional query-time step
- Harder to spam

HITS/Clever
- Query-dependent
- Per-query computation
- Small graph
- Outputs both hubs and authorities
- Easier to spam
- Quality depends on seed
Computational issues

- Web graph is
  - Huge
  - Sparse (average outdegree is under 20)
  - Changing

- Power iterations
  - Few iterations usually enough
  - Convergence of order good enough
Other issues

• Exploit structure [Arasu Novak Tomkins Tomlin 2001]
  – Degree distribution
  – Connectivity properties

• Stability questions
  – Stability/locality [Borodín Robers Rosenthal Tsaparas 2001]
  – [Ng Zheng Jordan 2001]

• Ranking the frontier
  – Ranking partially crawled pages [Eiron McCurley Tomlin 2004]
Monotonicity of PageRank

How does adding a new edge affect PageRank?

Theorem [Chien Dwork Kumar Simon Sivakumar 2002]:
Adding a new link to page $P$ can only
- Improve the PageRank value of $P$
- Improve the PageRank ordinal of $P$
Heuristic for incremental PageRank

- Locate the changed nodes
- Expand the seed set
- Recompute PageRank for expanded set
- Propagate values to the rest of the graph
Finding related pages

Input: One of more web pages
Goal: Find web pages that are related to input

Link-based algorithm [Dean Henzinger 1999]
  – Build a neighborhood graph around the input
  – Run HITS on the neighborhood graph
  – `Query-less’ mode
Building neighborhood graph

From p, go back, forward, back-forward, forward-back

Carefully limit the size of the graph
II: More search
Roadmap

• Duplicate detection
• Template elimination
• Link-based application: Focused crawling
• Link-based application: Trawling
• Link-based application: Web decay
• Search application: Intranet search
• Search application: WebFountain
Duplicate detection: Shingling

[Broder Glassman Manasse Zweig 1997]
Sketching/fingerprinting

D = domain of objects
f : D → R, a sketching function
   If f(a) ≠ f(b) then a ≠ b
   If a ≠ b then f(a) ≠ f(b) with high probability
   f(.) is easy and quick to compute

• Checking if two URLs are same
• Checking if two pages are near-duplicates
Why is this important?

• Page duplication
  – Mirrors (servers, documents/manuals)
  – Plagiarism
  – Minor modifications (email, last modified date, access counters, dynamic URLs)

• Expensive for crawling
• Expensive for indexing (memory, processing)
• 30% of web pages are duplicates
• Can be used to detect plagiarism
Shingle sets

Given a document $D = \{d_1, \ldots, d_n\}$, a $k$-shingle set $S_D$ is the (multi)-set of $k$-grams

Eg, $D = \{\text{Welcome to my homepage} \ldots\}$, $k = 2$

$S_D = \{\{\text{Welcome to}\}, \{\text{to my}\}, \{\text{my homepage}\}, \ldots\}$

Intuition: If $A$ and $B$ are near-duplicates then shingle sets overlap a lot, ie, $|S_A \cap S_B|$ is large
Jaccard coefficient

- Measure of intersection between two sets

\[ J(S_A, S_B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} \]

- \(1 - J(S_A, S_B)\) is a metric [Charikar 2002]
- \(J(S_A, S_B)\) large if A and B are near-duplicates
Min-wise independent permutations

Method to quickly test if $J(S_A, S_B)$ is large

$S_A, S_B \subseteq U$

$\pi: U \rightarrow U$, a random permutation

$a' = \min \{ \pi(a) \mid a \in S_A \}$

$b' = \min \{ \pi(b) \mid b \in S_B \}$

Min-wise lemma: $\Pr_{\pi}[a' = b'] = J(S_A, S_B)$
Proof of min-wise lemma

\[ c' = \min \{ \pi(c) \mid c \in S_A \cup S_B \} \]
\[ a' = b' \iff c' \in S_A \cap S_B \]

Probability this happens
\[ = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} = J(S_A, S_B) \]
Shingling algorithm

• Shingle sketch of a document
  – Minimal elements under $\pi_1(S_A)$, $\pi_2(S_A)$, ...
  – Expectation preserved
  – Variance is reduced
  – Truly random permutation not needed
    2-universal hashes work fine!
Template elimination

[Bar-Yossef Rajagopalan 2002]
Recall: Hypertext IR principles

- Relevant linkage principle
  \( p \) links to \( q \) \( \Rightarrow \) \( q \) is relevant to \( p \)

- Topical unity principle
  \( q_1 \) and \( q_2 \) are co-cited in \( p \) \( \Rightarrow \) \( q_1 \) and \( q_2 \) are related to each other

- Lexical affinity principle
  Closer the links to \( q_1 \) and \( q_2 \) are the stronger the relation between them

Fact: All these principles are systematically and frequently violated!
Violations of lexical affinity principle

- Alphabetical index lists/hubs
- HTML representation

Adjacent cells in the same column are far from each other in the HTML text
Templates

Template – Master HTML shell page used for composing new pages
Templates are bad for Web IR

• Violate the hypertext IR principles
  – Relevant linkage principle
  – Topical unity principle
• Extremely common
  – Web authoring tools
  – Standard in website design

Fact: Template elimination is crucial for effective Web search
Pagelets
[Chakrabarti 2001]

Pagelet – a region in a page that:

- Has a single theme
- Not nested within a bigger region with the same theme

Use pagelets for template elimination
Issues with pagelets

• **How to divide a page into pagelets?**
  – Use HTML cues
  – Machine learning

• **May lose semantic information in pages**
  – Use pages and pagelets together

• **No natural link structure on pagelets**
  – Pagelets point to pages

• **Adapt algorithms to work with pagelets**
  – HITS/Clever/PageRank can work with pagelets
Template elimination

• Characterizing properties of templates
  – Common look and feel
  – Controlled by a single authority

• A template is a collection $p_1, \ldots, p_k$ of pagelets satisfying
  – Similarity: $p_1, \ldots, p_k$ are identical or almost identical
  – Proximity: $p_1, \ldots, p_k$ belong to pages that are controlled by the same authority (eg, same website)

• Use shingling for similarity
Link-based application: Focused crawling

[Chakrabarti van den Berg Dom 1999]
Focused crawling

- Obtain pages within a specific topic
  - Web is huge
  - Full-scale crawling involves huge resources
    Disk space, bandwidth, crawling/processing time
  - Construct a focused portal (eg, bicycling)
  - Maintain high quality

Assumption: Relevant linkage principle, ie, p links to q ⇒ q is relevant to p
Naïve approach

- Fetch a page
- Check to see if it `belongs’ to a topic
- If so, retain
- If not, discard

Intuition: Visit and retain as many relevant pages and as few irrelevant pages as possible
Focused crawler: Construction

Crawler

Classifier

Visit priorities

Distiller

Relevance judgements

Seed
Seed set

- Given a topic, obtain example pages
  - Yahoo/OpenDirectory
  - Eg, if topic is bicycling, Yahoo! nodes could be bicyling, bicycle manufacturers, biking trails, ...
  - Topic is in a hierarchy

- Examples define topic, not the query
Distiller

• Identify hubs using HITS/Clever
• Use relevance of documents for weighting the edge
• Good hubs are crawled earlier
Link-based application: Trawling

[Kumar Raghavan Rajagopalan Tomkins 1999]
Communities on the web

• Where are the communities
  – Popular communities are listed
    • Yahoo!, OpenDirectory
    • Webrings, blogs (livejournal, xanga, orkut)
    • News groups, email lists
  – Subtler ones evolve/implicit

• Why study
  – Web sociology
  – Information organization
  – Marketing/commercial potential

• Thesis: Latent communities on the web outnumber the explicit ones by an order of magnitude
Link-based definition

- Communities = dense-bipartite subgraphs

- Why?
  - Insights from HITS
  - Links usually imply interest in a topic
  - Co-citation

Diagram showing a network of connections between entities like AT&T, MCI, Sprint, Alice, and Bob.
Communities from cores

- Finding communities not easy
- Core = small, complete bipartite subgraph

![Diagram of fans and centers]

- Fact: Every `large` enough `dense` bipartite graph `almost surely` has `small` core (eg, large = 3 x 10, dense = 50% edges, almost surely = 90% chance, small = 3 x 3)
Approach

• Preprocess the data

• Find all cores

• Expand cores into communities
Preprocessing

- Alexa crawl (1Tb data, 200M pages)
- Duplicate elimination
  - Syntactic (URL)
  - Content (Shingling)
- Popular page elimination
  - Don’t want too ‘popular’ communities
  - Popular community gives popular page
  - Popular page: indegree $\geq 50$
- Potential fans = has $> 5$ non-nepotistic links
Finding cores

- Database solution
  Find all triples of pages such that intersection of their outlinks is at least 3? Too expensive
- Eigen computations? Voluminous

- Heuristics
  - Pruning (Simple, inclusion-exclusion)
Simple pruning

- Examine each page if it is a potential fan or center
- Repeat
- Reduces to a sequence of sorting operations
Inclusion-exclusion pruning

a is a (3,*)-core if and only if the intersection of inlinks of x, y, and z is at least 3

- Include or exclude each page
- With index, can do it in main memory!
- Without index, needs two passes (only index of edges out of fans with indegree $\geq 3$)
Sample cores (out of 200K)

- Hotels in Costa Rica
- Clipart
- Japanese elementary schools
- Turkish student associations
- Oil spills off the coast of Japan
- Australian fire brigades
- Aviation/aircraft vendors
- Guitar manufacturers

Many of them were not present in 1999 Yahoo!
From cores to communities

- Use finding related-pages algorithm
  - Fans are exemplary hubs
  - Centers are exemplary authorities
  - Queryless operation

- Fossils can be recovered
Costa Rican hotels and travel

- The Costa Rica Inte...ion on arts, busi...
- Informatica Interna...rvices in Costa Rica
- Cocos Island Research Center
- Aero Costa Rica
- Hotel Tilawa - Home Page
- COSTA RICA BY INTER@MERICA
- tamarindo.com
- Costa Rica
- New Page 5
- The Costa Rica Internet Directory.
- Costa Rica, Zarpe Travel and Casa Maria
- Si Como No Resort Hotels & Villas
- Apartotel El Sestec... de San José, Cos...
- Spanish Abroad, Inc. Home Page
- Costa Rica's Pura V...ry - Reservation ...
- YELLOW\RESPALDO\HOTELES\Orquide1
- Costa Rica - Summary Profile
- COST RICA, MANUEL A...EPOS: VILLA
- Hotels and Travel in Costa Rica
- Nosara Hotels & Res...els &
- Restaurants...
- Costa Rica Travel, Tourism &
- Resorts
- Association Cívica de Nosara
- Untitled:
  http://www...ca/hotels/mimos.html
- Costa Rica, Healthy...t Pura Vida
- Domestic & International Airline
- HOTELES / HOTELS - COSTA RICA
- tourgems
- Hotel Tilawa - Links
- Costa Rica Hotels T...On line
- Reservations
- Yellow pages Costa ...Rica Export
- INFOHUB Costa Rica Travel Guide
- Hotel Parador, Manuel Antonio, Costa Rica
- Destinations
Japanese elementary schools

- The American School in Japan
- The Link Page
- Kids’ Space
- KEIMEI GAKUEN Home Page (Japanese)
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- ****E%o1•s—$’â‘ŠZ,J/fy[/fW
- Untitled: http://www...iglobe.ne.jp/~IKESAN
- fukui haruyama-es Home Page
- Torisu primary school
- goo
- Yakumo Elementary,Hokkaido,Japan
- FUZOKU Home Page
- Kamishibun Elementary School...
- schools
- LINK Page-13
- “ú…ţ ŠZ
- 100 Schools Home Pages (English)
- K-12 from Japan 10...net and Education)
- Untitled: http://www...iglobe.ne.jp/~IKESAN
- oŠ…‘ ŠZ,
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Home page (Japanese)
- UNIVERSITY
- %o—ŠZ DRAGON97-TOPO
- %’ê⁄4Å© YáYÉYá’¼ YáYÉYá’¼
Link-based application: Web decay

[Bar-Yossef Broder Kumar Tomkins 2004]
The changing web

- Web changes everyday
  - Average half-life of a page is quite short (few days)
  - Web littered with dead links
  - Changes are not predictable
- How to define quality wrt the changes?
- How do we know a page is not up-to-date?
  - Last modified date
  - Topics are quite out-dated
  - Dead links!
Automatically detecting decay

• Dead links
  – Easiest, noisy
• Last modified date
  – Server provided, not reliable
• Dates in the text
  – Difficult, noisy
• Understanding the text
  – Futuristic

Conclusion: Decay is hard to detect automatically
Definition of decay

Random surfer model (like PageRank)

• If current page is dead, output 1 (dead state)
• If current page is alive
  – W/p α, output 0 (alive state)
  – W/p 1 - α, choose a random outlink of current page and recurse

Decay(P) = probability that a random surfer starting at page P ends up in a dead state
Interpretation

- P is dead $\Rightarrow$ decay($P$) = 1
- P is alive, no links $\Rightarrow$ decay($P$) = 0
- P is alive, all links dead, $\Rightarrow$ decay($P$) = 1 - $\alpha$
- Decay($P$) = fraction of dead pages reachable from P, exponentially weighted by distance
Computing and using decay

- Computing decay
  - Recursive computation on the web graph
  - Can be approximated by random walks

- Using decay
  - Ranking web pages
  - Crawling decisions
  - Web sociology/economics
  - Web graph models
Decay scores for Yahoo!

Decay scores of pages from 30 Yahoo! nodes
Search application: Intranet search

[Fagin Kumar McCurley Novak Sivakumar Tomlin Williamson 2003]
Intranet vs. internet

- Different structure
- Democratic vs. autocratic or bureaucratic
- Approval processes, censorship, etc.
- Personal and organizational incentives
- Few hubs, but ultimate authorities
  - Not obvious if link-based search is effective
- Little published research
Axioms about intranets

1. Intranet documents created for simple dissemination of information, rather than to attract and hold attention
2. Large portions are not search-friendly
3. Intranets are essentially free of deliberate spam
4. Many queries tend to have a small set of correct answers (often unique!) and these pages are not easily identified
Potential ranking factors

• Anchortext vs. content vs. titles & metadata
• Query terms in the URL
• Length of URL (shorter is better)
• Depth of URL (fewer slashes are better)
• Hyperlink indegree (more is better)
• Length of document (more is better)
• Order of crawling (earlier is better)
• PageRank
Combining rankings...

U

Content index
Title index
Anchortext index
PageRank
Indegree
Discovery date
Words in URL
URL length
URL depth
Discriminator

TFxIDF

Result

Aggregation
**Intranet search: Lessons**

- Intranet search different from Internet search
  - Queries different (heavy on jargons, acronyms, etc)
  - Notion of good answer different (context-sensitive, user-sensitive)
  - Social processes of content creation different
- Efficacy of anchortext and title-keyword indices
- Customization at various levels very useful
  - Intranet-specific, user-specific, and query-specific
- Rank aggregation valuable tool for intranet search
Search application: WebFountain

IBM Almaden Research Center
Data Sources

Content Acquisition
- Crawler
- Ingest
- CO Mining

Data Analysis
- **Store** - Holds billions of pages of content and mined information.
- **CO Miners** - Analysis performed on every page.
- **Index** - Provides fast lookup to page content and mined data.
- **Dynamic Re-Index**
- **Entity Re-mine**
- **Entity Remine** – Customers request tracking of key terms through the corpus

Applications
- **WF and our partners deliver hosted applications based on:**
  - Temporal databases
  - On-topic stores
  - Customer applications that execute against the cluster

Cross-page cluster
- Trend DB
- On-topic Stores

Data is gathered by a large-scale crawler and by a number of data feeds provided by partners.
- Feeds include syndicated content, bulletin boards, weblogs, netnews, and data from customers.
Layers of Mining

Domain Specific Applications

EXTENSION MINERS
- Trend Analysis
- Dossier Creation
- Buzz Analytics
- Content Augmentation
- Brand Attributes
- Complex Queries

CROSS-PAGE MINERS
- Classification
- Clustering
- Associations, Seq. Patterns
- Ranking (Clever)
- Similarity
- Relationships

PER-PAGE MINERS
- Text Analyzer
- Links
- Regex
- People, Place, Date, etc.
- Geo Spatial
- SPAM, Porn, Dups
- Extraction, Tables, Lists

Declarative API

Data acquisition
- Web
- Intranet
- DB

Data Store
- Raw Data
- Meta Data
- Indexes
- Mining Results
III: Metasearch
Roadmap

- Metasearch problem and rank aggregation
- Voting and social choice
- Kemeny optimal/approximate aggregation
- Algorithms/heuristics and results
- Median rank aggregation and implications
- Other approaches to metasearch
Metasearch

MySearchEngine

Google
Yahoo!
msn
alltheweb
WiseNut
TEOMA
Why metasearch?

• **Coverage:** Search engines don’t overlap much
• **Consensus ranking:** Get the best out of several ranking heuristics
• **Spam resistance:** Hard to fool many search engines
• **Query robustness:** Work for both broad-topic and specific queries
• **Feedback:** Reflects the effectiveness of a particular search engine
Combining ranking functions

Aggregate ranking

Links

Page title

Anchor text

URL

Last modified date

Text
Similarity search in databases

Given collection of \( n \) database elements (each is a \( d \)-tuple of attributes) and given at run-time a query element \( q \) (another \( d \)-tuple of attributes) find the database element that best matches \( q \)

Each of the \( d \) attributes is a voter

Database elements = candidates

Each voter ranks all candidates

Database elements ranked by voter \( i \), based on similarity to the query \( q \) in attribute \( i \)

Find top winners of this election by aggregation
Basic theme: Rank aggregation

Input: n candidates and k voters

 Preferential voting: Each voter gives a (partial) list of the candidates in order of preference

```
1 3     ...     10
3 19    ...     17
7 n     ...     1
...
n 10    ...
```

Goal: Produce a good consensus ordering of all n candidates

Deja vu: Voting/elections
Voting/elections

- Politics, jury decisions, pooling expert opinions, program committees, …
- More than balance subjective opinions
  Seek the truth
  Find the “best” candidate, second “best”, …
- What is “best”?
- Majority opinion represents (objectively) best?
CS vs SC

- Small number of voters
- Large number of candidates
- Algorithmic efficiency
- Input could be partial lists/top $k$ lists
- Limited overlap among top results
- Output might have to be a ranking
Desiderata (CS)

- Simple algorithm
- Fast algorithm (near-linear time)
- Provable quality of solution
- If approximation, factor should be independent of number of candidates/voters
Borda’s proposal (1770)

Election by order of merit

First place is worth 1 point, second place is worth 2 points ...

Candidate’s score = Sum of points

Borda winner: Lowest scoring candidate

Eg, MVP in MLB
Condorcet’s proposal (1785)

Partition candidates into $A, B$

If for every $a \in A$ and $b \in B$, a majority ranks $a$ ahead of then aggregation must place all elements in $A$ ahead of all elements in $B$

Condorcet winner: A candidate who defeats every other candidate in pairwise majority-rule election

Marie J. A. N. Caritat, Marquis de Condorcet
Condorcet ≠ Borda

Borda scores: A \((1*6 + 3*4 = 18)\), B \((2*6 + 1*4 = 16)\), C \((3*6 + 2*4 = 26)\)

B is the Borda winner

Condorcet criterion: A beat both B and C in pair-wise majority

A is the Condorcet winner
Condorcet paradox

A  B  C
B  C  A
C  A  B

Condorcet winner may not exist!

Black (1950s): Choose Condorcet winner; if none, choose Borda winner
Copeland (1951): Choose candidate with highest outdegree – indegree in the majority graph
Many other voting schemes

- **Plurality vote**
  - Candidate with most # first positions is winner

- **Instant runoff vote**
  - President of Ireland, Australian parliament, many US university student elections

- **Single-transferable vote**
  - Malta, Republic of Ireland, Australian Senate

- ...
Arrow’s theorem (1951)

The following are irreconcilable

• Every result must be achievable somehow
• Monotonicity: Ranking higher should not hurt a candidate
• Independence of irrelevant attributes: Changes in rankings of “irrelevant alternatives” should have no impact on ranking of “relevant” subset
• Non-dictatorship

Conclusion: ⌈ satisfactory rank aggregation function
Borda vs. Condorcet debate

- **Borda**
  - Score-based
  - Consistent: two separate set of voters yield same ranking $\Rightarrow$ their union yields same ranking
  - Any score-based method not Condorcet

- **Condorcet**
  - Majority-based
  - Meet Arrow’s criteria where “independence of irrelevant attributes” criterion is modified
  - Winner may not exist
Kemeny’s proposal (1959)

Axiomatic approach

- “Distance” between two preference orderings
  \[ \text{Distance} = \text{number of pair-wise disagreements} \]
- Obtain ordering that is “least-distant” from the individual orderings

Theorem [Young Levenglick 1988]: Kemeny’s rule is the unique preference function that is neutral, consistent, and Condorcet

- Reconciles Borda and Condorcet
- Satisfies additional properties (Pareto, anonymity)
- Maximum likelihood interpretation: [Young 1988]
Metrics on permutations

- Domain: \([n] = \{1, 2, \ldots, n\}\)
- \(\sigma \in S_n\)
- \(\sigma(i) < \sigma(j)\) means that “\(\sigma\) ranks \(i\) above \(j\)”

Kendall \(\tau\) distance

Spearman’s footrule distance
Kendall $\tau$ distance

$K(\sigma, \tau) = \text{Number of pairs } (i, j) \text{ such that } \sigma \text{ ranks } (i, j) \text{ in one order and } \tau \text{ ranks them in the opposite order}$

- Bubble-sort distance
- $K$ is a metric
- $K$ is right invariant: $K(\sigma, \tau) = K(\sigma \tau^{-1}, 1)$
- Eg

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>number of disagreements: 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

$(AB, AD, CD)$
Spearman’s footrule distance

\[ F(\sigma, \tau) = \sum_{i = 1, n} |\sigma(i) - \tau(i)| \]

- \( F \) is a metric (L₁ norm)
- \( F \) is right invariant: \( F(\sigma, \tau) = F(\sigma \tau^{-1}, 1) \)
- \( \text{Eg.,} \)

\[
\begin{array}{cc}
A & B \\
B & D \\
C & A \\
D & C \\
\end{array}
\]

shift(A) = 2  shift(B) = 1, etc., so

footrule distance: 6
There are several others, but...

Many of the other metrics are computationally expensive (some NP-hard, some not known to be polynomial-time computable, etc.)

[Diaconis; Group Representation in Probability and Statistics]

Also these two are perhaps the most natural for many applications
Diaconis--Graham inequality

\[ K(\sigma, \tau) \leq F(\sigma, \tau) \leq 2 \, K(\sigma, \tau) \]

This inequality is essentially tight
\[ F(\sigma) \leq 2 \, K(\sigma) \]

\[
F(\sigma) = \sum_i |\sigma(i) - i| \\
= \sum_i \left| \sum_j [\sigma(i) > \sigma(j)] - [i > j] \right| \\
\leq \sum_i \sum_j |[\sigma(i) > \sigma(j)] - [i > j]| \\
= \sum_{i,j} [\sigma(i) > \sigma(j), i < j] \\
= 2 \, K(\sigma)
\]
\[ K(\sigma) \leq F(\sigma) \]

- \([i: j] = \text{inversion } i < j, \sigma(i) > \sigma(j)\]
  - Type 1 inversion: if \(\sigma(i) \geq j \Rightarrow i < j \leq \sigma(i)\)
    \[ \Rightarrow \forall i, \#\{ j \mid [i; j] \text{ is type 1 inversion} \} \leq \sigma(i) - i \]
  - Type 2 inversion: if \(\sigma(i) \leq j \Rightarrow \sigma(j) < \sigma(i) \leq j\)
    \[ \Rightarrow \forall j, \#\{ i \mid [i; j] \text{ is type 1 inversion} \} \leq j - \sigma(j) \]

- Every inversion is type 1, or type 2, or both

\[ K(\sigma) \leq \text{type 1 inversion + type 2 inversion} \]
\[ \leq \sum_{i \mid \sigma(i) > i} (\sigma(i) - i) + \sum_{j \mid j > \sigma(j)} (j - \sigma(j)) \]
\[ \leq F(\sigma) \]
Optimal rank aggregation

Given metric $d(\cdot, \cdot)$ and input permutations $\sigma_1, \ldots, \sigma_k$, find permutation $\pi^*$ such that

$$\sum_{i=1}^{k} d(\sigma_i, \pi^*)$$

is minimized

Kemeny (Kendall) optimal aggregation: $d = K$

Spearman footrule optimal aggregation: $d = F$
Kemeny optimal aggregation

Theorem [Bartholdi Tovey Trick 1989]: Kemeny optimal aggregation is NP-hard

Theorem [DKNS]: Kemeny optimal aggregation is NP-hard even for 4 lists

– Reduction using feedback edge set
c-approximate aggregation

Given metric \( d(\cdot, \cdot) \) and input permutations \( \sigma_1, \ldots, \sigma_k \), find permutation \( \pi \) such that

\[
\sum_{i=1}^{k} d(\sigma_i, \pi) \leq c \cdot \sum_{i=1}^{k} d(\sigma_i, \pi^*)
\]
Trivial approximation

Theorem: $2(1 - 1/k)$-approximation can be computed easily

Proof: $K$, $F$ are metrics and simple geometry

$\pi^* = \text{Optimal aggregation wrt. } d(\cdot, \cdot)$

$i^* = \arg \min_i \sum_j d(\sigma_i, \sigma_j)$

$\sum_j d(\sigma_j, \sigma_{i^*}) \leq (1/k) \sum_{j, j'} d(\sigma_j, \sigma_{j'})$

$\leq (1/k) \sum_{j, j'} (d(\sigma_j, \pi^*) + d(\pi^*, \sigma_{j'}))$

$\leq 2 \sum_j d(\sigma_j, \pi^*)$
Footrule optimal aggregation

Theorem [DKNS]: F-optimal aggregation can be computed in polynomial time

Proof: Via minimum cost perfect matching

\[
\sum_{i = 1, k} |\sigma_i(a) - p|\]
2-approximation to K-optimum

Use Diaconis--Graham inequality

$\pi = $ Footrule optimal aggregation

$\pi^* = $ Kendall-optimal aggregation

\[
\sum_i K(\sigma_i, \pi) \leq \sum_i F(\sigma_i, \pi) \\
\leq \sum_i F(\sigma_i, \pi^*) \\
\leq 2 \sum_i K(\sigma_i, \pi^*)
\]

Open question: Better factor approximations for Kemeny optimum? Hardness?
Heuristics: Markov chains

- States = candidates
- Transitions = function of preference orders
  Probabilistically switch to a better candidate
- Final ranking = order of stationary probabilities
Advantages of Markov chains

- Handling partial lists and top $k$ lists using available information to infer new ones
- Handling uneven comparisons and list lengths
- Motivation from PageRank---more wins better, more wins against good players even better
- With $O(nk)$ preprocessing, $O(k)$ per step for about $O(n)$ steps
Sample Markov chains

If current state is candidate P, next state is:

• **MC1:** Choose uniformly from the multiset of all candidates that were ranked higher than or equal to P by some voter that ranked P

...  

• **MC4:** Choose uniformly a candidate Q from all candidates and switch if the majority preferred Q to P
Metasearch results

- Using top 100 from AV, AW, EX, GG, HB, LY, NL
- Queries: affirmative action, alcoholism, ...

<table>
<thead>
<tr>
<th></th>
<th>K</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borda</td>
<td>0.214</td>
<td>0.345</td>
</tr>
<tr>
<td>Footrule</td>
<td>0.111</td>
<td>0.167</td>
</tr>
<tr>
<td>MC1</td>
<td>0.130</td>
<td>0.213</td>
</tr>
<tr>
<td>MC2</td>
<td>0.128</td>
<td>0.210</td>
</tr>
<tr>
<td>MC3</td>
<td>0.114</td>
<td>0.183</td>
</tr>
<tr>
<td>MC4</td>
<td>0.104</td>
<td>0.149</td>
</tr>
</tbody>
</table>
Heuristics: Median

Theorem [DKNS]: If the median ranks of the candidates are unique (ie, form a permutation), then this permutation is a footrule optimal aggregation.

What about using the median itself for ranking, even if it is not unique?
Median rank aggregation

Given \( \sigma_1, \ldots, \sigma_k \),

\[ \mu'(i) = \text{median} \ (\sigma_1(i), \ldots, \sigma_k(i)) \]

Order \( \mu' \) to obtain a permutation \( \mu \)

Eg,

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>A</td>
</tr>
</tbody>
</table>

\( \mu'(A) = 3, \ \mu'(B) = 2, \ \mu'(C) = 3, \ \mu'(D) = 2 \)

\( \mu = B \ D \ A \ C \)
Median is a good approximation

Theorem $[^{FKMSV}]$: Median rank aggregation is a 3-approximation to footrule optimal aggregation

Median ranking is used in Olympic figure skating

Open question: Is the constant 3 tight?
Consistent permutations

Given $\sigma' = \sigma'_1, \ldots, \sigma'_n$ where $\sigma'_i \in R$, call a
permutation $\sigma \in S_n$ to be consistent with $\sigma'$ if
$$\sigma'_i < \sigma'_j \Rightarrow \sigma(i) < \sigma(j)$$

Consistency lemma: If $\sigma$ is consistent with $\sigma'$, then
for any other permutation $\tau$, $F(\sigma, \sigma') \leq F(\tau, \sigma')$
Proof of consistency lemma

Fact: \(a' \leq b'\) and \(a < b\) \(\Rightarrow\)
\[
|a - a'| + |b - b'| \leq |a - b'| + |a' - b|
\]

If \(\tau \neq \sigma\), apply this fact repeatedly to differing pairs until \(\tau\) becomes \(\sigma\)
Each time \(F(\tau, \sigma')\) can only improve
Median lemma

Fact: Given $x_1, \ldots, x_n$ where $x_i \in \mathbb{R}$,

\[
\text{median}(x_1, \ldots, x_n) = \arg \min_y \sum_i |x_i - y|
\]

Median lemma: Given permutations $\sigma_1, \ldots, \sigma_k$, let $\mu'$ denote their median function. Then, for any permutation $\tau$,

\[
\sum_i F(\mu', \sigma_i) \leq \sum_i F(\tau, \sigma_i)
\]
Proof of median theorem

Let $\tau$ be any permutation

$$\sum_i F(\mu, \sigma_i) \leq \sum_i F(\mu, \mu') + \sum_i F(\mu', \sigma_i) \quad \text{(triangle)}$$

$$\leq \sum_i F(\tau, \mu') + \sum_i F(\mu', \sigma_i) \quad \text{(consistency)}$$

$$\leq \sum_i F(\tau, \sigma_i) + 2 \sum_i F(\mu', \sigma_i) \quad \text{(triangle)}$$

$$\leq \sum_i F(\tau, \sigma_i) + 2 \sum_i F(\tau, \sigma_i) \quad \text{(median)}$$

$$= 3 \sum_i F(\tau, \sigma_i)$$
Merits of median

- Simple to implement
- Admits instance optimal algorithms: among all algorithms that do sequential and random access to pre-sorted preference orders, the run-time of this median-finding algorithm is optimal up to a factor of 2
- Provably good method for nearest-neighbor applications
Borda rank aggregation

Given \( \sigma_1, \ldots, \sigma_k \),

\[
\beta'(i) = \sigma_1(i) + \cdots + \sigma_k(i)
\]

Order \( \beta' \) to obtain a permutation \( \beta \)

Eg,

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>A</td>
</tr>
</tbody>
</table>

\( \beta'(A) = 8, \beta'(B) = 6, \beta'(C) = 8, \beta'(D) = 8 \)

\( \beta = B A C D \)
Borda is a good approximation

Theorem [FKMSV]: Borda rank aggregation is a 5-approximation to footrule optimal aggregation

Borda lemma: \[ \sum_i F(\beta', \sigma_i) \leq 2 \sum_i F(\mu', \sigma_i) \]
Prove this point-wise for every \( j \) in the domain

Open question: Aggregating wrt other metrics on permutations (eg, Borda is near-optimal wrt Spearman’s rho)
Copeland rank aggregation

Given $\sigma_1, \ldots, \sigma_k$,

$$\Gamma(i, j) = \text{majority}\ \{\sigma_1(i) \text{ vs } \sigma_1(j), \ \sigma_k(i) \text{ vs. } \sigma_k(j)\}$$

$$\gamma'(i) = \sum_i \Gamma(i, j) - \sum_j \Gamma(j, i)$$

Order $\gamma'$ to obtain a permutation $\gamma$

Eg,

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>B</td>
<td>D</td>
<td></td>
<td>D</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td></td>
<td>A</td>
</tr>
</tbody>
</table>

$\gamma'(A) = -1, \ \gamma'(B) = 3, \ \gamma'(C) = -1, \ \gamma'(D) = -1$

$\gamma = B \ A \ C \ D$
Copeland is a good approximation

Theorem [FKMSV]: Copeland rank aggregation is a 6-approximation to Kendall optimal aggregation

Proof: As before, but using K instead of F
Plurality method

Given $\sigma_1, \ldots, \sigma_k$,

$$\pi'(i) = \langle \ldots, \# j\text{-th place votes, }\ldots \rangle$$

Lexicographically order $\pi'$ to obtain a permutation $\pi$

Eg,

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>D</td>
<td></td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\pi'(A) = \langle 1 \ 0 \ 1 \ 1 \rangle$, $\pi'(B) = \langle 1 \ 1 \ 1 \ 0 \rangle$,

$\pi'(C) = \langle 1 \ 0 \ 1 \ 1 \rangle$, $\pi'(D) = \langle 0 \ 2 \ 0 \ 1 \rangle$

$\pi = B \ A \ C \ D$
Plurality is not a good approximation

Theorem [FKMSV]: Plurality rank aggregation is not a good to approximation to Kendall optimal aggregation

Proof: n candidates, k voters, $n \gg k$

1 1 2 3 4 ... k-1 $\pi = 1 2 \ldots n$
2 2 3 2 2 ... 2 $\sum_i F(\pi, \sigma_i) \geq (k-2)(n-1)$
3 3 4 4 3 ... 3 $\beta = 2 3 \ldots n 1$

$\sum_i F(\beta, \sigma_i) \leq k^3 + n$

n n 1 1 1 ... 1 $n \uparrow \Rightarrow \text{Ratio} = \Omega(k)$
More rank aggregation applications

- Comparing search engine quality [DKNS, FKS]
- Spam reduction [DKNS]
- Intranet search [FKMNSTW]
- Similarity search [FKS]
- Multiple-criteria selection (eg, travel, restaurant)
- Word association techniques (AND queries) [DKNS]
Other approaches to Metasearch

• **Support vector machines** [Joachims 2002]

• **Learning** [Cohen Schapire Singer 1999]
  – Hedge algorithm—iterative weight update

• **Condorcet fusion** [Montague Aslam 2002]
  – Finding Hamiltonian paths in Condorcet graphs

• **Bayesian** [Aslam Montague 2001]