Information Retrieval on the Web

Massimo Melucci
melo@dei.unipd.it

Outline of Second Part

- Algorithms for Information Retrieval on the Web
- Evaluation of Information Retrieval on the Web

Algorithms for Information Retrieval on the Web

Outline

- Use of Web links for IR on the Web
- Hypertext Structuring, Analysis, and Metrics
- Web Page Authority Analysis
- Automatic Web Link Generation

Use of Web links for IR on the Web

- Web IR system activities are performed automatically and involve the design and implementation of algorithms and data structures.
  We concentrate on the search engine, and not on the search agent.
- The main difference between the algorithms for IR on the Web and the ones for classical IR is the massive presence of Web links.
  Web links are source of evidence and of noise.
- Web links represent a relationship between the connected pages.
- Research in IR has been focussing on semantic relationships:
  the Cluster Hypothesis, for example.

Topics Related to the Use of Web Links

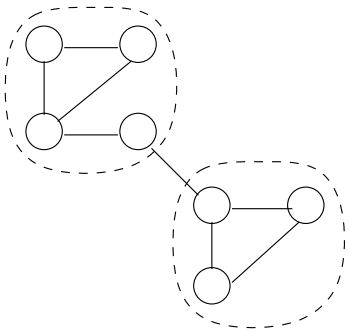
- The idea of using relationship representations, such links or citations is not completely new:
  It is usual in IR;
- the notion of impact factor as a means to assess the importance of a scientific journal;
- the notion of Web impact factor to assess the role of Web pages and Web sites;
- hypertext analysis, structuring and metrics;
- bibliographic citation analysis for information retrieval and document clustering.

Hypertext Structuring, Analysis, and Metrics

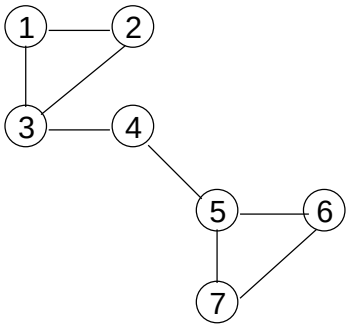
1. hypertext structuring – clustering nodes using content or links;
2. hypertext analysis – detecting node roles, e.g. index or reference;
3. hypertext metrics – to measure hypertext properties;

Clustering Nodes

- clustering nodes using *content* is usually done in IR; the clustered nodes are textual documents
  e.g., see the van Rijsbergen's or Salton's textbooks
- within hypertext, *links* are used to cluster nodes: For example, the higher the number of paths between two nodes, the more the nodes are close.
  Links express semantic closeness.
  See Botafogo *et al.*



Detecting Node Roles



	1	2	3	4	5	6	7	O_i
1	–	1	1	∞	∞	∞	∞	5.7
2	1	–	1	∞	∞	∞	∞	5.7
3	1	1	–	1	∞	∞	∞	6.3
4	∞	∞	1	–	1	∞	∞	5.7
5	∞	∞	∞	1	–	1	1	6.3
6	∞	∞	∞	∞	1	–	1	5.7
7	∞	∞	∞	∞	1	1	–	5.7
I_j	5.7	5.7	6.3	5.7	6.3	5.7	5.7	

After replacing ∞ with a constant, say 2:

$D = \sum_i \sum_j M_{ij} = 57 \quad O_i = \frac{D}{\sum_j M_{ij}} \qquad I_j = \frac{D}{\sum_i M_{ij}}$

Nodes 3 and 5 can be index nodes, e.g. homes or roots, because their out-centrality (O_i), which is equal to the in-centrality (I_j) because of simmetry, is high.

Example of Metrics to Measure Hypertext

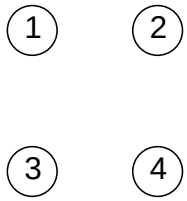
- developed to capture the notions of connectedness and complexity of hypertext;
- *compactness*: average number of in-links and out-links; *stratum*: average number of links to follow to reach a node:
  useful for "hierarchical" hypertexts, e.g. Web sites;
- compactness and stratum are related to some evaluation measures:
  - many nodes can be reached if the hypertext is compact; if navigation is for IR, recall can be high, but precision can be low;
  - navigation effort to reach nodes is low if the hypertext is little stratified.

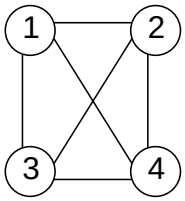
Example of Compactness

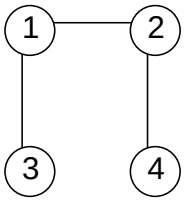
	1	2	3	4
1	0	K	K	K
2	K	0	K	K
3	K	K	0	K
4	K	K	K	0

	1	2	3	4
1	0	1	1	1
2	1	0	1	1
3	1	1	0	1
4	1	1	1	0

	1	2	3	4
1	0	1	1	2
2	1	0	2	1
3	1	2	0	3
4	2	1	3	0







$MAX = n(n-1)K$ $\quad$ $MIN = n(n-1)$ $\quad$ $Cp = \frac{MAX - \sum_i \sum_j M_{ij}}{MAX - MIN}$

if $K = 4$: $Cp = 0$ $\quad$ $Cp = 1$ $\quad$ $Cp = \frac{48-20}{48-12} = 0.78$

Use of Hypertext Analysis, Structuring and Metrics for IR on the Web

- there are some fundamental differences between the Web and other information systems that make Web hypertext analysis important;
- the Web is much more complex than a local hypertext or a bibliographic database:
  high heterogeneity, untyped links, highly dynamic;
- the Web is the result of the convergence, or the divergence, of a myriad of contributions from, and is accessed by million of end user;
- the knowledge about Web link topology is incomplete;
  only for Web samples, which must be processed to detect all the out- and in-links

Use of Hypertext Analysis, Structuring and Metrics for IR on the Web

- combination of link- and content-based clustering methods:
  link-based methods are independent of the content, so they can help overcome some keyword mismatch problems;
  content-based methods can discover relationships that are not coded by Web links (see automatic Web link generation);
- structural analysis, e.g. hierarchy discovering, help detecting Web link types;
  for example, links of a hierarchy may represent specialization or aggregation relationships;

Use of Hypertext Analysis, Structuring and Metrics for IR on the Web

- clusters and other structures, like hierarchies, can be indexed, retrieved and displayed as an individual object;
  for example, metadata of root nodes, e.g. home pages, can be used to index children nodes; clusters can be presented to be further examined;
- metrics can help discover hidden structures, guide the generation of clusters and the selection of hierarchies;
  for example, select clusters with a given internal connectivity, select hierarchies with a given out-centrality;

Two Algorithms for Web Page Authority Analysis Techniques

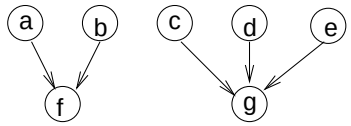
- HITS (Hyperlink Induced Topic Search): focusses on broad topic queries that are likely to be answered with too many pages;
- PageRank (Google): simulates a random walk across the Web and compute the score of a page as the probability of reaching the page;
- the effectiveness of these algorithms is based on the assumption that:
  - the more the page is pointed to by other pages, the more the page is popular;
  - popular pages are more likely to include relevant information than non-popular pages.

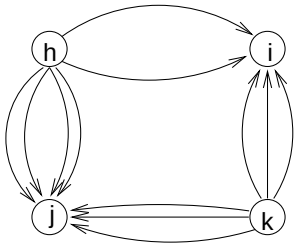
HITS (Hyperlink Induced Topic Search)

- broad topic queries produce large retrieved page sets;
- these sets are likely to include popular pages;
- some popular pages are authoritative pages;
- some authoritative pages are relevant pages;

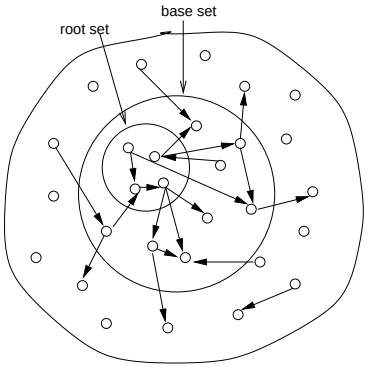
HITS (Hyperlink Induced Topic Search)

- two types of page: hub and authority;
- circular relationship:
  "good" hubs point to "good" authorities,
  "good" authorities point to "good" hubs;





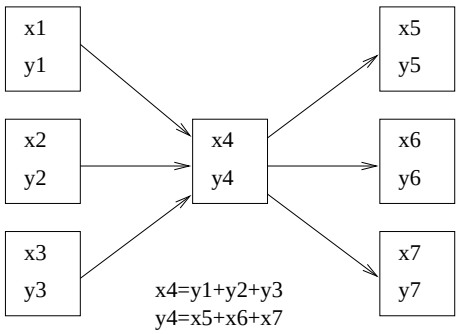
HITS: the algorithm



q: a query
R_q: the root set retrieved to answer q
B_q: the base set after expanding R_q using links
A_q: the authorities
H_q: the hubs
S_q: the final result set
k: a natural number
σ: a threshold

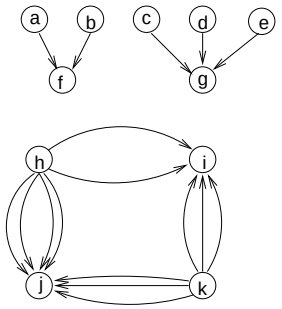
for each q
  R_q = answer(q)
  B_q = expand(R_q)
  (A_q, H_q) = iterate(B_q, k)
  S_q = filter(A_q, H_q, σ)
end for

HITS: Updating Authority and Hub Weights



iterate(B_q, k)
B_q: the base set with cardinality $|B_q|$
k: the number of iterations
z: the vector $(1, ..., 1) \in \mathcal{R}^{|B_q|}$
$x_{(i)}$: authority weight vector at step i
$y_{(i)}$: hub weight vector at step i
$x_{(0)} = z$
$y_{(0)} = z$
for each $i = 1, ..., k$
    $x_{(i)}$ = update($y_{(i-1)}$);
    $y_{(i)}$ = update($x_{(i)}$);
    normalize $x_{(i)}$ and $y_{(i)}$
end for
return $(x_{(k)}, y_{(k)})$

HITS: Updating Authority and Hub Weights



	$x_{(k)}, y_{(k)}$	
	$x_{(0)}, y_{(0)}$	$x_{(1)}, y_{(1)}$
a	1,1	0,0.5
b	1,1	0,0.5
c	1,1	0,0.5
d	1,1	0,0.5
e	1,1	0,0.5
f	1,1	0.6,0
g	1,1	0.8,0
h	1,1	0,0.7
i	1,1	0.6,0
j	1,1	0.8,0
k	1,1	0,0.7

for example:
$x_{(1)}^{(f)} = y_{(0)}^{(a)} + y_{(0)}^{(b)}$
$y_{(1)}^{(c)} = x_{(1)}^{(g)}$
then, normalize x, y using $||x||, ||y||$, respectively, where $||z|| = \sqrt{\sum_i z_i^2}$

PageRank

- designed by Brin and Page at Stanford University;
- they implemented Google, a search engine based on PageRank;
- PageRank rationale: A page has high rank if the sum of the ranks of its back-links is high;
  back-link of page p: a link from a page to p, i.e. in-link
  forward-link of page p: a link from p to a page, i.e. out-link
  a high PageRank page has many back-links or few highly ranked back-links

PageRank Intuitive Definition

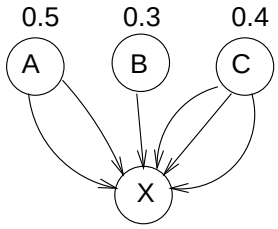
- a random surfer clicks on links at random;
  "back" button is not used;
- if the surfer gets into a loop of Web pages, the surfer will jump to some other page;
- the surfer jumps to a page with probability d, called "damping factor".

PageRank Formal Definition

- let $Pr(X)$ be the probability that the surfer is at page X at a given time;
- let $Y_1, ..., Y_n$ be the n pages pointing to X, and $C_i, C_i > 0, i = 1, ..., n$ be the number of out-links from Y_i to X;
- then, the page rank of X is given by the following recursive expression:

$$Pr(X) = K(1-d) + Kd(\frac{Pr(Y_1)}{C_1} + \cdots + \frac{Pr(Y_n)}{C_n})$$

where K is a normalization factor



$$Pr(X) = (1 - 0.15) + 0.15(\frac{0.3}{2} + \frac{0.3}{1} + \frac{0.4}{3})$$

PageRank vs. HITS

- the PageRank is computed for all the Web pages stored in the database and then prior to every query; HITS is performed on the set of retrieved Web pages, and then for each query;
- HITS computes authorities and hubs, while PageRank computes authorities only;
- the implementation details of PageRank has been reported.

Some Problems with HITS and PageRank

- *mutual reinforcement* because of the recursive move of weights between the same hubs and authorities;
  pages of site A always and only cite pages of site B and viceversa;
- *ambiguous links* occur because different authors cite a page for different reasons, which may contrast one to each other;
- *tool generated links* cannot be easily detected and are treated equally to manually generated links;
  e.g., links generated by search engines;
- *topic drift*, or change, causes these algorithms to produce hubs and authorities about specializations or generalizations of the query topic;

Modifications to HITS and PageRank

- *weighing authorities and hubs* $(w_A(h,j), w_H(h,j))$ help deal with mutual reinforcement;
  avoiding intra-server links helps as well;
- *weighing links* $(S(h,j))$ help deal with topic drift;

update(v)
v': new vector weight
for each $j = 1, ..., |B_q|$
$v'_j \leftarrow \sum_h v_h \times w_A(h,j) \times w_H(h,j) \times S(h,j) \times I(h,j)$
where $I(h,j) = 1$ if a link
exists between pages h and j
$I(h,j) = 0$ otherwise
$w_A(h,j)$ is the authority weight
$w_H(h,j)$ is the hub weight
$S(h,j)$ is the similarity
between pages h and j
end for
return v'

Some Remarks

- PageRank and HITS may reinforce authorities and hubs:
  users cite authorities and hubs in their Web pages, the new Web pages increases authority and "hubbiness";
- non-authority or non-hub pages are missed, yet relevant:
  authority is not relevance;
- PageRank and HITS process a sample of the Web at a given time period: results may change if a different sample is drawn;
  relationships with search agents;
- these algorithms process links equivalently, but not all links are equal:
  some sort of link filtering should be useful;
  automatic link construction algorithms may be used.

Automatic Web Link Generation

- IR has been dealing with detecting, creating and managing links between informative objects;
- associative links are those of interest for IR because they:
  1. express semantic relationships, and
  2. are the necessary means for navigation and browsing;
- efficiency reasons: manual link generation is infeasible for usual document collection;
- effectiveness reasons: IR deals with semantic content-based retrieval, then links should express semantic content-based relationships;
- automatic link generation methods are then necessary, because of efficiency reasons, and useful, because of effectiveness reasons.

Algorithms for Automatic Web Link Generation

- Web links should be characterized, e.g. "typed", filtered, or weighted, to reach effective IR on the Web;
- some links are for navigational purposes, others are advertisement links, others represent semantic relationships that can be exploited;
- algorithms for automatic Web link generation are useful to detect new links, assign a type to, or filter links;

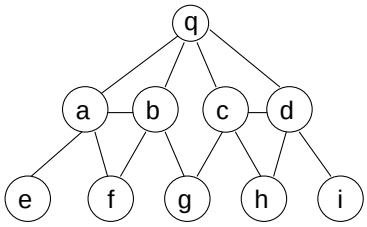
Methods for Automatic Link Generation

Basic Method	Methods
1. compute the descriptions of two textual node content using a model, e.g. the vector-space model; 2. compute the similarity score between the descriptions using a similarity function, e.g. the cosine; 3. if the function value is over the given threshold, then insert the link between the nodes.	• the use of text similarity to generate links; • the automatic detection of links and of different link types; • the automatic construction of links of different types between different types of node.

The Use of Text Similarity to Generate Links

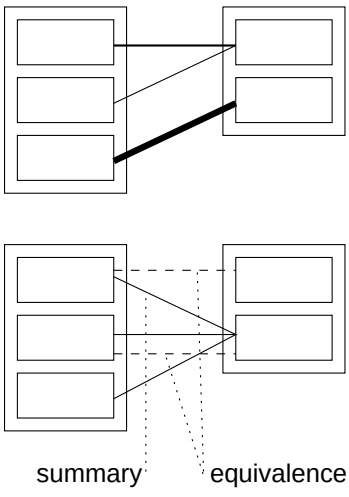
retrieve(x, n, k)
if($n > 0$) then
    extract top k text segments matching x
    for each segment $y_i, i = 1, ..., k$
        retrieve($y_i, n - 1, k$)
end retrieve

Let n, k be natural numbers and q be the starting query; then, retrieve(q, n, k);



The Automatic Detection of Links of Different Link Types

1. compute similarity links between text segments;
2. collapse high-similarity (strong) links and merge linked segments;
3. refine low-similarity (weak) links and split linked segments;
4. detect link types:
   revision, summary, expansion, equivalence, contrast, tangent, aggregate;

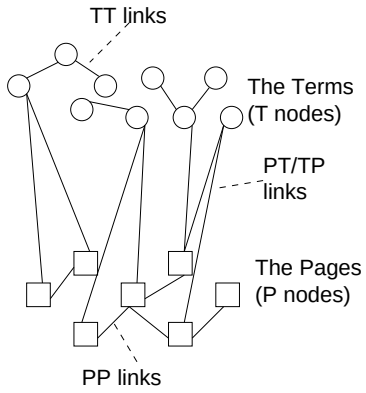


Extracting Text Segments from Web Pages

- research in text structuring and passage retrieval has dealt with newspaper article, encyclopaedia articles, or TREC documents;
- the consistency between text structure and text content affects the effectiveness of the developed methods;
  e.g., a paragraph is about a subject and the subject is organized in (continuous) paragrahs;
- HTML may help extracting text segments, but:
  - segments may be very short and meaningless, i.e. they carry few evidence about the subject;
  - extracting segments requires parsers that are able to detect HTML page anomalies, e.g. missing tags, repeated `BODY` tag;

The Automatic Construction of Links of Different Types between Different Types of Node

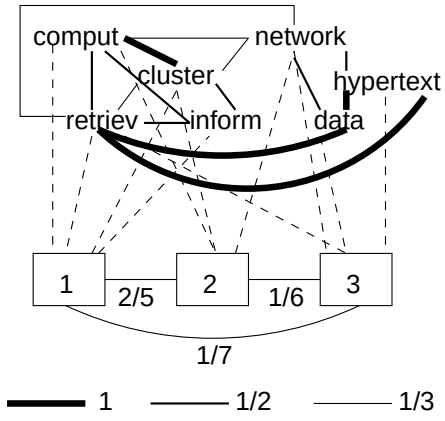
- PP (TT) links are between pages (auxiliary data); the link weight is the cosine of the angle between the page (auxiliary data) vectors;
- PT links are placed between a page and auxiliary data, while TP links are placed between an auxiliary datum and pages. the weight, say $tf \times idf$, of an auxiliary datum within a page can be used as numerical measure of link strength.



Example of a Two Levels Hypertext

1 The Computation of Clustering for Information Retrieval
2 Clusters of Computer Networks
3 Data Retrieval with Hypertextual Networks

Links are computed and weighed using $\frac{|X \cap Y|}{|X \cup Y|}$, where X, Y are either sets of keyword stems or sets of documents.



Evaluation of Information Retrieval on the Web

Outline

- Some Issues Evaluation of IR on the Web
- The Cranfield Model and Evaluation of IR on the Web
- Web Link Navigation Measures
- The Web Track of TREC

Evaluation of Information Retrieval on the Web

- operational and laboratory experiments;
  - operational – e.g., study of real Web searchers using search engines and tools;
  - laboratory – e.g., use of test collections accordingly to the Cranfield model
- the object of evaluation – the whole Web, a search engine as a "black box", or a specific model, technique, method.

Some Issues Evaluation of IR on the Web

- some issues characterize evaluation of IR on the Web, and affect specifically the Cranfield model;
- *dynamicity* of the Web and of search engines;
- *heterogeneity* of pages and queries;
- *hyperlinking* among Web pages.

The Cranfield Model and Evaluation of IR on the Web

- representativeness of a test collection;
  - the Web changes very rapidly, is highly heterogeneous, Web pages are linked, users perform different tasks;
    e.g., CGI script-generated pages;
  - test documents should be representative of the Web in terms of heterogeneity, hyperlinks, time span;
  - test queries should also be for question answering, resource finding, etc.;
- notion of relevance;
  - Web pages are linked directly,
  - users do necessarily see pages one after another, and
  - their judgements about the relevance of a page does strongly depend on the judgements given on the previously seen pages;
  - relevance may change because different browsers display different content;
    e.g., Lynx displays the NOFRAMES part, while Navigator displays the frame content;

The Cranfield Model and Evaluation of IR on the Web

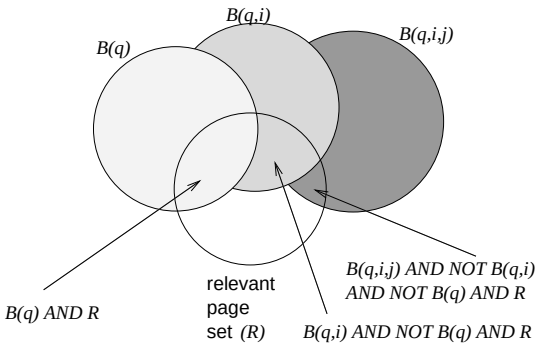
- notion of relevance (cont.);

	Linked Page	
Linking Page	relevant	non-relevant
relevant	relevant,useful	relevant
non-relevant	useful	non-relevant, useless

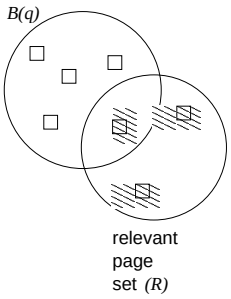
- measures of effectiveness;
  - computing absolute recall is impossible;
  - recall computation using recall for very specific topics or using pooling method;
  - navigation measures about the variations of recall and precision.

Two Web Link Navigation Measures

- novelty – the increase of relevant pages retrieved at a given navigation step;
- noise – the increase of non-relevant pages retrieved at a given navigation step:

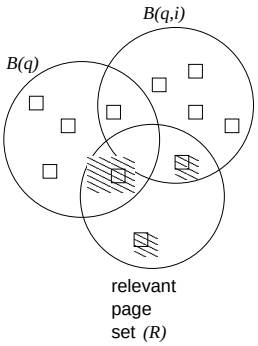


An Example of Computation of Novelty and Noise



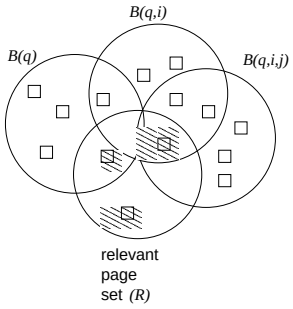
Novelty = Recall = $\frac{1}{3}$

Noise = 1-Precision = $\frac{4}{5}$



Novelty = $\frac{1}{3}$

Noise = $\frac{4}{6}$



Novelty = 0

Noise = $\frac{3}{6}$

The Web Track of TREC

- aims to provide experimental results about the performance of IR on the Web within a scientific standard framework;
- based on the Cranfield model;
- at TREC-8 (1999), two main tasks – the Small Web task and the Large Web task;
- the Small Web task: 2GB of Web data, 250,000 documents; The Large Web task: 100GB of Web data, 18.5 million documents;
- objectives:
  - comparison between Web track and Ad-Hoc track;
  - effectiveness of new algorithms, e.g. HITS and PageRank;

Some Results from the Web track

- context, visual rendition and position of anchors may be important to determine the relevance of linked pages;
  anchors placed on top the page, rendered with color, larger sizes, or meaningful text
- the notion of relevance should be modified;
- need of evaluating link-based algorithms;
- Web queries are very often different from TREC topics (e.g. size, typology)