QuieW
Quality-based View of the Web

P.Avesani, D.Sona
ITC-irst
Web Directory

http://www.dmoz.org
SegnaWeb: risorse Internet selezionate dai bibliotecari italiani

Affari e economia
- Commercio
- Consumi e utenze
- Economia, imprese e management
- Finanza, moneta e banche
- Fisco
- Lavoro e concorsi
- Pensioni e sicurezza sociale
- Pesi e misure
- Statistiche
- Trasporti e traffico

Arte e cultura
- Architettura
- Arti visive e plastiche
- Cinema
- Collezionismo
- Fotografia
- Fumetti
- Letteratura italiana
- Letterature
- Libri e editoria
- Moda
- Musei, gallerie e mostre
- Musica
- Spettacoli e teatro
- Televisione e radio
- Urbanistica e paesaggio

Attualità e media
- Ecologia, meteo e ambiente
- Emergenze
- Eventi
- Appuntamenti nazionali
- Giornali e agenzie di stampa
- Lotto, lotterie, premi
- Ona osata
- Sport
- Televisione e radio

Divertimento e tempo libero
- Animali domestici
- Auto e moto
- Casa, giardino e bricolage
- Cucina
- Feste, sagre e folklore
- Giochi e hobbies
- Lotto, lotterie, premi
- Sport
- Televisione e radio
- Turismo e viaggi

Informatica e Internet
- E-mail gratuita
- Grafica e immagini
- Informatica
- Internet
- Motori di ricerca e directory
- Software gratuito

Istruzione e formazione
Renewal

- **Quality-based Sampling**
  Manual selection of web sources

- **Ranking Seeds**
  Dealing with Google Bombs

- **Shallow Ontologies**
  Extensional definition of concepts

- **Context-driven Queries**
  Focused search
Renewal

**Quality-based Sampling**
Manual selection of web sources

**Ranking Seeds**
Dealing with Google Bombs

**Shallow Ontologies**
Extensional definition of concepts

**Context-driven Queries**
Focused search
The bubble of web visibility

Marco Gori* and Ian Witten*

* Dipartimento di Ingegneria dell’Informazione, Università di Siena
* Department of Computer Science, University of Waikato, New Zealand

The web seems like a Borgesian library with a huge amount of information. Access to this treasure is mediated by dragons who compete amongst themselves for dominance. The battleground is so hostile that few will survive; indeed, just one will likely achieve overall dominance in all but specialist corners of the library. The allegory is attractive except that the treasure is not private property but a public good, and the dragons must serve their own business interests rather than the public cause to compete successfully.

A vital source of information in everyday life, all the content of the web is readily available in principle. But the dominant mode of access is through search engines, and—in practice—the view they offer is restricted. Typical queries match hundreds of thousands of documents, which search engines sort according to ranking heuristics. For pragmatic reasons users view only the highest ranked pages, which gives rise to the notion of web page “visibility” as seen through the lens provided by search engines.

The heuristics adopted by popular search engines favor the construction of artificial communities that are expressly

Figure 1: How to promote a web page by playing on the number of input links. A keyword-page is created which links to a set of artificial pages containing the selected keywords which, in turn, link to the target page.
Renewal

- **Quality-based Sampling**
  Manual selection of web sources

- **Ranking Seeds**
  Dealing with Google Bombs

- **Shallow Ontologies**
  Extensional definition of concepts

- **Context-driven Queries**
  Focused search
Renewal

- **Quality-based Sampling**
  Manual selection of web sources

- **Ranking Seeds**
  Dealing with Google Bombs

- **Shallow Ontologies**
  Extensional definition of concepts

- **Context-driven Queries**
  Focused search
Query

http://www.acm.org/sigs/sigkdd/kdd2005

KDD-Cup 2005

News

August 30, 2005: KDD-Cup presentation slides from the KDD conference
August 30, 2005: Winning Teams
August 30, 2005: Labeled Query Data
August 10, 2005: Solution Evaluation Result

Introduction

The KDD-Cup 2005 Knowledge Discovery and Data Mining competition will be held in conjunction with the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The task is selected to be interesting to participants from both academia and industry. In particular, we encourage the participation of students. This year's competition is about classifying Internet user search queries. We are looking forward to an interesting competition and encourage your participation.

Contest Rules

Agreement

By sending the registration email, you indicate your full and unconditional agreement and acceptance of these contest rules.

Eligibility

The contest is open to any party planning to attend KDD 2005. A person can participate in only one group. Multiple submissions per group are allowed, since we will not provide feedback at the time of submission. Only the last submission before the deadline will be evaluated and all other submissions will be discarded.
Assumptions

**Paper versus Web Sources**
- Present: classification of knowledge on paper
- Future: classification of knowledge on internet

**Representation versus classification**
- Extensional encoding of categories
Assumptions

**Paper versus Web Sources**
- Present: classification of knowledge on paper
- Future: classification of knowledge on internet

**Representation versus classification**
- Extensional encoding of categories
Outline

Motivation
- Web directories revisited

Technology
- Web directories back-stage

Research
- Web directories intelligent supports
Technology
Platform
<table>
<thead>
<tr>
<th>Technology</th>
<th>Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Directory</td>
<td>Taxonomy of categories</td>
</tr>
<tr>
<td>Tagging</td>
<td>Anytime tags association</td>
</tr>
<tr>
<td>Bookmarking</td>
<td>Manual on-the-fly selection</td>
</tr>
<tr>
<td>Searching</td>
<td>Full text focussed search</td>
</tr>
<tr>
<td>Crawling</td>
<td>Monitoring of web sources</td>
</tr>
<tr>
<td>Classification</td>
<td>Qualifying web sources</td>
</tr>
<tr>
<td>Ranking</td>
<td>Topic-driven relevance criteria</td>
</tr>
<tr>
<td>Recommendation</td>
<td>Hints on source classification</td>
</tr>
<tr>
<td>Feeds Aggregation</td>
<td>Organization of RSS feeds</td>
</tr>
<tr>
<td>Reviewing</td>
<td>Annotations of web resources</td>
</tr>
</tbody>
</table>
### Technology Prototype

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web Directory</strong></td>
<td>Taxonomy of categories</td>
</tr>
<tr>
<td><strong>Tagging</strong></td>
<td>Anytime tags association</td>
</tr>
<tr>
<td><strong>Bookmarking</strong></td>
<td>Manual on-the-fly selection</td>
</tr>
<tr>
<td><strong>Searching</strong></td>
<td>Full text focused search</td>
</tr>
<tr>
<td><strong>Crawling</strong></td>
<td>Monitoring of web sources</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td>Qualifying web sources</td>
</tr>
<tr>
<td><strong>Ranking</strong></td>
<td>Topic-driven relevance criteria</td>
</tr>
<tr>
<td><strong>Recommendation</strong></td>
<td>Hints on source classification</td>
</tr>
<tr>
<td><strong>Feeds Aggregation</strong></td>
<td>Organization of RSS feeds</td>
</tr>
<tr>
<td><strong>Reviewing</strong></td>
<td>Annotations of web resources</td>
</tr>
</tbody>
</table>

- **Top-Down Strategy**
- **Bottom-up Strategy**
<table>
<thead>
<tr>
<th>Technology Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web Directory</strong></td>
</tr>
<tr>
<td>Taxonomy of categories</td>
</tr>
<tr>
<td><strong>Tagging</strong></td>
</tr>
<tr>
<td>Anytime tags association</td>
</tr>
<tr>
<td><strong>Bookmarking</strong></td>
</tr>
<tr>
<td>Manual on-the-fly selection</td>
</tr>
<tr>
<td><strong>Searching</strong></td>
</tr>
<tr>
<td>Full text focussed search</td>
</tr>
<tr>
<td><strong>Crawling</strong></td>
</tr>
<tr>
<td>Monitoring of web sources</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
</tr>
<tr>
<td>Qualifying web sources</td>
</tr>
<tr>
<td><strong>Recommendation</strong></td>
</tr>
<tr>
<td>Hints on source classification</td>
</tr>
<tr>
<td><strong>Feeds Aggregation</strong></td>
</tr>
<tr>
<td>Organization of RSS feeds</td>
</tr>
<tr>
<td><strong>Reviewing</strong></td>
</tr>
<tr>
<td>Annotations of web resources</td>
</tr>
</tbody>
</table>

Static Perspective **versus** Dynamic Perspective
Outline

Motivation
- Web directories revisited

Technology
- Web directories back-stage

Research
- Web directories intelligent supports
Research

Taxonomy-driven Classification
- Learning categorization of web sources

Category-driven Ranking
- Learning custom relevance criteria

Sharing Lexicon
- Distributed coordination on tagging annotation
Research

Classification
Thanks

Questions?

Acknowledgment: we are here thanks to the Geek Librarian :-)
Project MeTe
“Metodi e percorsi per ricerche tematiche su internet”
Partnership Liceo Tambosi
Grant Fondazione Caritro 2006-2007
We have designed the graphical user interface of QuiW prototype. It is conceived as a web application where to manage the discovery, the filtering, the classification, the organization, the ranking and the tagging of web resources.

Feel free to have a look at the full size snapshot clicking on the thumbnail above. Please leave a comment below on your first reaction (it has been proved that it is enough few milliseconds to recognize a user friendly interface).

Posted in Case Studies, Tools | Comments

More