Databases and Digital Libraries

Faculty

Bill Arms

Alan Demers

Johannes Gehrke

Jayavel Shanmugasundaram

Researchers

Mirek Riedewald

Carl Lagoze
Current Projects

- **Cougar (Sensor Databases)**
  - Demers, Gehrke
- **eScience (Scientific Data Management)**
  - Arms, Demers, Gehrke, Ridewald, Shanmugasundaram
- **Himalaya (Data Mining)**
  - Gehrke, Riedewald
- **Privacy in Databases**
  - Gehrke
- **Quark (Unifying Databases and IR)**
  - Shanmugasundaram
- **NSDL (Digital Libraries)**
  - Arms, Lagoze
Keyword Search over HTML

Query Keywords ➔ Google™ ➔ Ranked Results

Hyperlinked HTML Documents
Car for sale! I have a great car for sale. It is a 1995 Ford Mustang convertible. Excellent condition (except for a little rust). $2000 or best offer. Please click here for a picture.
<workshop date="28 July 2000">
  <title>XML and Information Retrieval: A SIGIR 2000 Workshop</title>
  <editors>David Carmel, Yoelle Maarek, Aya Soffer</editors>
  <proceedings>
    <paper id="1">
      <title>XQL and Proximal Nodes</title>
      <author>Ricardo Baeza-Yates</author>
      <author>Gonzalo Navarro</author>
      <abstract>We consider the recently proposed language …</abstract>
      <section name="Introduction">
        Searching on structured text is becoming more important with XML …
        <subsection name="Related Work">
          The XQL language …
        </subsection>
      </section>
      …
      <cite xmlns:xlink="http://www.acm.org/www8/paper/xmlql">…</cite>
    </paper>
    …
  </proceedings>
</workshop>
Keyword Search over XML

Query Keywords → XRank → Ranked Results

Mix of Hyperlinked XML and HTML Documents
XML and Information Retrieval: A SIGIR 2000 Workshop

David Carmel, Yoelle Maarek, Aya Soffer

Proceedings

Paper ID 1

Title: XQL and Proximal Nodes

Authors: Ricardo Baeza-Yates, Gonzalo Navarro

Abstract: We consider the recently proposed language ...

Section: Introduction

Searching on structured text is becoming more important with XML ...

Subsection: Related Work

The XQL language ...

Cite: http://www.acm.org/www8/paper/xmlql
<workshop date="28 July 2000">
<title>XML and Information Retrieval: A SIGIR 2000 Workshop</title>
<editors>David Carmel, Yoelle Maarek, Aya Soffer</editors>
<precedings>
<paper id="1">
<title>XQL and Proximal Nodes</title>
<author>Ricardo Baeza-Yates</author>
<author>Gonzalo Navarro</author>
<abstract>We consider the recently proposed language …</abstract>
<section name="Introduction">
Searching on structured text is becoming more important with XML …
</section>
<subsection name="Related Work">
The XQL language …
</subsection>
</paper>
</precedings>
</workshop>
Design Principles

1) Return most specific element containing the query keywords
<workshop date="28 July 2000">
  <title>XML and Information Retrieval: A SIGIR 2000 Workshop</title>
  <editors>David Carmel, Yoelle Maarek, Aya Soffer</editors>
  <proceedings>
    <paper id="1">
      <title>XQL and Proximal Nodes</title>
      <author>Ricardo Baeza-Yates</author>
      <author>Gonzalo Navarro</author>
      <abstract>We consider the recently proposed language …</abstract>
      <section name="Introduction">
        Searching on structured text is becoming more important with XML …
        <subsection name="Related Work">
          The XQL language …
        </subsection>
      </section>
    </paper>
    <paper id="2">
      ...
    </paper>
  </proceedings>
</workshop>
Design Principles

1) Return most specific element containing the query keywords
2) Ranking has to be done at the granularity of elements
<workshop date="28 July 2000">
  <title>XML and Information Retrieval: A SIGIR 2000 Workshop</title>
  <editors>David Carmel, Yoelle Maarek, Aya Soffer</editors>
  <proceedings>
    <paper id="1">
      <title>XQL and Proximal Nodes</title>
      <author>Ricardo Baeza-Yates</author>
      <author>Gonzalo Navarro</author>
      <abstract>We consider the recently proposed language …</abstract>
      <section name="Introduction">
        Searching on structured text is becoming more important with XML …
        <subsection name="Related Work">
          The XQL language …
        </subsection>
      </section>
    </paper>
  </proceedings>
</workshop>
Design Principles

1) Return most specific element containing the query keywords
2) Ranking has to be done at the granularity of elements
3) Generalize HTML keyword search
Design Principles

1) Return most specific element containing the query keywords

2) Ranking has to be done at the granularity of elements

3) Generalize HTML keyword search
Data Model


Containment edge

Hyperlink edge
ElemRank

• Objective importance of element
• Analogous to Google’s PageRank
  – But computed at granularity of elements
  – Exploit hyperlink edges and containment edges
• Naturally generalizes Google’s PageRank
  – Random walk interpretation
PageRank [Brin & Page 1998]

\[
p(v) = d \times \sum_{(u,v) \in HE} \frac{p(u)}{N_h(u)} + \frac{1 - d}{Nd}
\]

- d: Probability of following hyperlink
- 1-d: Probability of random jump
- Hyperlink edge

Diagram: Nodes and edges representing the PageRank formula.
**ElemRank**

- **Hyperlink edge**: \( d_{1/3} \)
- **Containment edge**: \( d_{2/2} \)

\[ e(v) = d_1 \sum_{(u,v) \in HE} \frac{e(u)}{N_h(u)} + d_2 \sum_{(u,v) \in CE} \frac{e(u)}{N_c(u)} + d_3 \sum_{(u,v) \in CE^{-1}} e(u) + \frac{1 - d_1 - d_2 - d_3}{N(v)} \]

- \( d_1 \): Probability of following hyperlink
- \( d_2 \): Probability of visiting a subelement
- \( d_3 \): Probability of visiting parent
- \( 1-d_1-d_2-d_3 \): Probability of random jump
Summary

• Broad research areas
  – Sensors, data mining, privacy, eScience, digital libraries

• Collaborative research
  – Systems, machine learning, security/privacy
  – Peta-byte data store for astronomy, graphics, web algorithms

• Vibrant group of talented students
  – 14 PhD students