CWS: A Comparative Web Search System

†Jian-Tao Sun, ‡Xuanhui Wang, §Dou Shen
†Hua-Jun Zeng, †Zheng Chen

† Microsoft Research Asia
‡ University of Illinois at Urbana-Champaign
§ Hong Kong University of Science and Technology
Problem to Solve

• Massive needs for comparing information
  – Products, stores, companies
  – Peoples, countries, cities
  – General information

• Few effective ways for this comparison
  – Existing comparison shopping engines (e.g. shopping.com and froogle.google.com)
    • Domain-dependent
    • Based on structured information
  – Search engines
    • Single search box
    • Long list of search result pages
A Scenario for Information Comparison

• Comparing **Greece** and **Turkey** for a holiday
• Method 1:
  – Input “Greece vs. Turkey” into a search engine
  – Some results with low quality
A Scenario for Information Comparison (cont’l)

• Method 2:
  – Input “Greece” and “Turkey” separately
  – Good results for each query but difficult to compare them

Simple result list problem
Our Proposal

• Comparative Web Search (CWS):
  – Facilitate the information comparison by using search engines

• Features:
  1. Multiple search boxes for the input
  2. Side-by-side comparison of corresponding results
  3. Clustering related results into themes
Related Work

• Website comparison [Liu, WWW02; Liu, KDD01]
  – Hierarchical clustering webpages of two websites
  – Pages are displayed as a tree form
  – Differences are highlighted

• Comparative Web browser [Nadamoto, WWW03]
  – Concurrently presents multiple Web pages
  – After a user selects a page from one site, the system retrieves similar contents from the other site
• Comparative text mining, [Zhai, KDD04; Zang, Master thesis 2004]
  – Mining a set of comparative text collections
  – Discover latent common themes and specific themes across all collections
• Product comparison [Hu, KDD04; Liu, WWW05]
  – Extract customers' opinions on product features based on a collection of customer reviews
  – Both customers and manufacturers can make comparisons between products
Figure 1: The Flowchart of CWS System.
CWS Pair-view Interface

1. Canon Sure Shot 130U 35mm Film Camera - Find, Compare, and Buy at...
   Read Reviews and Compare Prices on Canon Sure Shot 130U 35mm Film Camera. DealTime helps shoppers find, compare, and buy anything in just seconds.
   http://www.dealtyme.com/PC-Canon_Sure_Shot_130U

2. Canon Sure Shot 130u Reviews
  won, Canon Sure Shot 130u Reviews - PhotographyReview.com is the leading resource of quality consumer-generated product reviews on the Internet.

3. Canon Sure Shot 130u - Point & Shoot / Zoom camera - 35mmprices...
   CNET Shopper.com - compare prices and availability for Canon Sure Shot 130u - Point & Shoot / Zoom camera - 35mm from our quality stores.
   http://shopper.cnet.com/Canon_Sure_Shoot_130u_Point_Shoot_Zoom_camera_35mm/2014.6503.9.30254584.html

4. Canon Sure Shot 130U - Reviews, Best Prices and Product...
   BizRate helps you buy the Canon Sure Shot 130U for the lowest price by comparing Canon Sure Shot 130U prices at the top-rated stores online.

5. Compare Prices and Read Reviews on Canon Sure Shot 130U 35mm Film...
   Epinions has the best comparison shopping information on Canon Sure Shot 130U 35mm Film Camera. Compare prices from across the web and read reviews from...

6. Olympus Stylus Epic QD 35mm Film Camera - Find, Compare, and Buy...
   Olympus Stylus Epic QD 35mm Film Camera - Find, Compare, and Buy...
   http://www.dealtyme.com/PC-Olympus_Stylus_Epic_QD

7. Olympus Stylus Epic QD Reviews
   Olympus Stylus Epic QD Reviews - Olympus Stylus Epic QD Reviews - PhotographyReview.com is the leading resource of quality consumer-generated product reviews on the Internet.

8. Olympus Stylus Epic QD - Point & Shoot camera - 35mmprices - CNET...
   CNET Shopper.com - compare prices and availability for Olympus Stylus Epic QD - Point & Shoot camera - 35mm from our quality stores.
   http://shopper.cnet.com/Olympus_Stylus_Epic_QD_Point_Shoot_camera_35mmfrom/our-quality/197.prod_id-6786487.html
CWS Cluster-view Interface

Common keywords for clusters

Query-specific keywords
Algorithm for Page Pair Ranking

- Input: query q₁ & q₂
- Output: ranked list of comparative page pairs
- Assumptions: page pair <p₁, p₂> is a comparative page pair if:
  - p₁ is relevant to q₁
  - p₂ is relevant to q₂
  - <p₁, p₂> contains comparative information of q₁ and q₂
Algorithm for Page Pair Ranking

- Function for measure the comparativeness of page pair \(<p_1, p_2>\)

\[
f_{q_1, q_2}(p_1, p_2) = \alpha \cdot R(p_1, q_1) + \beta \cdot R(p_2, q_2) + \lambda \cdot T_{q_1, q_2}(p_1, p_2)
\]

\[
T_{q_1, q_2}(p_1, p_2) = \theta \cdot S(\text{url}_1, \text{url}_2) + (1 - \theta) \cdot S(p_1 \setminus q_1, p_2 \setminus q_2)
\]

\(\forall p_1 \in SR_1, p_2 \in SR_2\)

- \(f\): Comparativeness function
- \(R\): Relevance between query and page
- \(S\): Similarity between two text segments
- \(SR\): Search result list
- \(T\): Comparative information contained in the page pair

\(p \cdot q\cdot\): Remaining text content of page \(p\) after removing \(q\) from it
Algorithm for Clustering and Keyword Extraction

- Cluster comparative page pairs
  - Each page pair \(<p_1, p_2>\) is treated as a whole
  - A probabilistic clustering algorithm based on simple mixture generative model [Zhai, KDD04]

\[
P(w|\theta_d) = (1 - \lambda_B) \sum_{j=1}^{k} [\pi_{d,j} P(w|\theta_j)] + \lambda_B P(w|\theta_B)
\]

\[
log P(C|\Omega) = \sum_{d \in C, w \in V} c(w, d) \cdot log P(w|\theta_d)
\]

\[
\hat{j} = \text{argmax}_j \{\pi_{d,j} | j = 1, \cdots, k\}
\]

- Represent clusters by keywords \(P(w|\theta_j)\)
Algorithm for Clustering and Keyword Extraction

• Extracting query-specific keywords
  ▪ Supervised keyword extraction algorithm
    ✓ Linear regression model with 4 features
      – PF: phrase frequency
      – ATF: average frequency of all terms in phrase
      – AIDF: average inverse document frequency
      – OKA: OKAPI weighting score

  ▪ Selection of key-phrases for sub-clusters
    ✓ Entropy based approach

\[ Ent(w) = - \sum_{i=1,2} p_i \log p_i \]
Experiment for Page Pair Ranking

• Data set
  ✓ 20 query pairs
  ✓ Retrieve top 50 pages of MSN search for each query

<table>
<thead>
<tr>
<th></th>
<th>$q_1$</th>
<th>$q_2$</th>
<th>$q_1$</th>
<th>$q_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>xbox</td>
<td>playstation</td>
<td>11</td>
<td>sars</td>
</tr>
<tr>
<td>2</td>
<td>Sony dv</td>
<td>Samsung dv</td>
<td>12</td>
<td>McDonalds</td>
</tr>
<tr>
<td>3</td>
<td>Canon sure shot 130u</td>
<td>Olympus stylus epic</td>
<td>13</td>
<td>Nike</td>
</tr>
<tr>
<td>4</td>
<td>lancome</td>
<td>clinique</td>
<td>14</td>
<td>Iraq war</td>
</tr>
<tr>
<td>5</td>
<td>Ford Escape</td>
<td>Jeep Liberty</td>
<td>15</td>
<td>virtual earth</td>
</tr>
<tr>
<td>6</td>
<td>PSP</td>
<td>GBA</td>
<td>16</td>
<td>Chengxiang Zhai</td>
</tr>
<tr>
<td>7</td>
<td>Maradona</td>
<td>Pele</td>
<td>17</td>
<td>Sony Camera</td>
</tr>
<tr>
<td>8</td>
<td>Nokia cell phone</td>
<td>Motorola cell phone</td>
<td>18</td>
<td>windows</td>
</tr>
<tr>
<td>9</td>
<td>MIT</td>
<td>CMU</td>
<td>19</td>
<td>MSN messenger</td>
</tr>
<tr>
<td>10</td>
<td>Microsoft</td>
<td>Google</td>
<td>20</td>
<td>Bush</td>
</tr>
</tbody>
</table>

Table 1: 20 Pairs of Comparative Queries
Data Labeling and Evaluation

- 3 human labelers judge the results of pair-view mode
  - Is the left page relevant with the first query?
  - Is the right page relevant with the second query?
  - Is the page pair helpful for making comparisons?

- Evaluation Method:
  - Precision@N
    - Number of correct comparative page pairs in top N / N
• Conclusions:
  – Our algorithm can get a 80% top1 precision
  – Both URL and snippet are useful for comparativeness measure
  – The combination of them get best result
Page Pair Ranking Case Study

• Comparative page pair examples
  ▪ “Canon Sure Shot 130u” vs. “Olympus Stylus Epic”
  ▪ “Afghanistan War” vs. “Iraq War”

   http://www.bartleby.com/65/af/AfghanWar.html

2. The Observer — Special reports — War in Afghanistan
   http://observer.guardian.co.uk/afghanistan/0,1501,522841,00.html

3. Afghanistan Timeline, 21st Century
   http://www.mapreport.com/countries/afghanistan.html

4. Articles about September 11 2001 attacks on USA and Afghanistan
   http://people.pwf.cam.ac.uk/nwm20/usa_afghanistan.html

5. War Report - Iraq War and Afghan Aftermath - con news
   http://www.conw.org/warreport/

6. Independent Online Edition > World Politics
   http://news.independent.co.uk/world/politics/article574313.ece

7. Government Resources
   http://library.louisville.edu/government/subjects/world/politics/article574313.ece

8. Amazon.com: The Lessons of Afghanistan: War Fighting

9. Afghanistan : War Without End?
   http://www.pbs.org/newshour/bb/asia/afghanistan/

10. Iraq
    Recent Additions

Iraq
Shiite Power Struggle Simmers in Najaf

U.S. to Intensify Its Training in Iraq to Battle Insurgents

Afghanistan
Recent Additions
CIA Holds Terror Suspects in Secret Prisons

Detainee Policy Sharply Divides Bush Officials

As Gitmo Hunger Strike Continues... Lawyers Step Up Fight for Access

‘Failure Is Not an Option’

Figure 4: A comparative page returned for query pair: ‘Afghanistan war’ and ‘Iraq war’.
Experiment for Comparative Page Clustering

- Example results of comparative page clustering and keyphrase extraction

Table 3: Keyphrase Extraction Result for Query Pair: $q_1 = \text{‘ChengXiang Zhai’}, q_2 = \text{‘Jiawei Han’}$

<table>
<thead>
<tr>
<th>Common Keyphrases</th>
<th>$q_1$ Specific Keyphrases</th>
<th>$q_2$ Specific Keyphrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. illinois, urbana, champaign (44)</td>
<td>university, filtering, collaborative</td>
<td>mellon university, list, pakdd-2001 tutorials</td>
</tr>
<tr>
<td>2. research, system, database (44)</td>
<td>beespace, automated, news-gazette online</td>
<td>mining, participation, concepts</td>
</tr>
<tr>
<td>3. author, title, resource (44)</td>
<td>annual, information retrieval, embedding</td>
<td>data mining, data, anhai</td>
</tr>
<tr>
<td>4. author, track, kdd (24)</td>
<td>information retrieval, research, anhai</td>
<td>mining, conference, data</td>
</tr>
<tr>
<td>5. usa, tao, award (26)</td>
<td>papers, zhai cs hong, zhang fa</td>
<td>diff, delete, business intelligence</td>
</tr>
</tbody>
</table>
Conclusions

• In this work
  ▪ Proposed and studied a new search problem, comparative Web search
  ▪ Implemented a CWS system, characterized by
    ✓ Allowing users input two comparative queries
    ✓ Organizing pages into ranked comparative page pairs
    ✓ Grouping page pairs into comparative clusters
    ✓ Extraction of keyphrases to summarize comparative information

• Future work
  ▪ Adoption of other evaluation approaches for larger scale experiment
  ▪ Automatic identification of comparative query pairs
Q&A

Microsoft Research Asia

hjzeng@microsoft.com
Backup Slides
<table>
<thead>
<tr>
<th>#</th>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Greece hotel, Greek hotel, Greek island, Greece travel, Greece</td>
<td>We provide information on Greece hotel, Greek hotel, Greek island, Greece travel, Greece tour, Greece vacation, Greece holiday. Greece accommodation, ... <a href="http://www.greethotel.com">http://www.greethotel.com</a></td>
</tr>
<tr>
<td>3</td>
<td>Greece Holidays: Cheap Holidays in Greece, Europe</td>
<td>Lastminute.com offers fantastic savings on Greece holidays lastminute.com, the first place to look for cheap Greece Holiday deals at the last minute. <a href="http://www.lastminute.com/site/find/World/Europe/Greece/Greece--Holiday.html">http://www.lastminute.com/site/find/World/Europe/Greece/Greece--Holiday.html</a></td>
</tr>
<tr>
<td>7</td>
<td>Turkey Holidays: Cheap Holidays in Turkey, Europe</td>
<td>Lastminute.com offers fantastic savings on Turkey holidays. lastminute.com, the first place to look for cheap Turkey Holiday deals at the last minute. <a href="http://www.lastminute.com/site/find/World/Europe/Turkey/Turkey--Holiday.html">http://www.lastminute.com/site/find/World/Europe/Turkey/Turkey--Holiday.html</a></td>
</tr>
</tbody>
</table>
### Comparative Web Search

Intelligently align two search queries’ result snippets to facilitate comparison.

#### Results Map

- Sony camera, canon camera (6)
  - powershot, cybershot (15)
  - comparing prices (10)
    - shopping products (5)
  - powershot (4)
  - cybershot (3)
  - accessories (3)
    - others (1)
  - others (6)
- str (6)
- memory cards (5)
- Olympus (5)
- Purchase (3)
- photography (5)
- Jsp (4)
- pixel (4)
- php forums (4)
- others (16)

#### Cluster compare prices contains 10 documents

1. **Sony Camera - Compare Prices, Reviews and Buy at NexTag - Price ...**
   
   Sony Camera: 215 results like the SONY Digital Camera 6.0 M Pixel 3x Optical Zoom, SONY CyberShot DSC-T9 Digital Camera, SONY DSC-H2 Digital Camera, and more...
   
   [Link](http://www.nextag.com/sony-camera/zzulic/biz20/search-html)

2. **Sony Camera Accessories: Read reviews and compare prices at Ciao ...**
   
   Searching for a camera accessory? Find the perfect camera case, converter, camera lens, data cable, strap and much more...
   
   [Link](http://www.ciao.co.uk/Camera_Accessories_5276379_2-Sony)

3. **Sony Digital Cameras - Compare Prices, Reviews and Online Stores ...**
   
   To buy the best Sony digital camera for you, compare specs, prices ... Sony Digital Camera Reviews, Sales on Sony Digital Cameras, Sony Camera Comparison...
   
   [Link](http://www.bizrate.com/digital_cameras/brand-sony/products__at259--390.html)

4. **Sony Digital Camera - Top 10 digital cameras**
   
   Buy Sony digital cameras from the best UK shops online. Click on one of the products listed below to compare prices to find the best price to order...

---

Canon Sd550 - Compare Prices, Reviews and Buy at NexTag - Price ...

Canon Sd550: 10 results like the Canon Powershot SD550 Digital Camera, Canon PowerShot SD550 7.1MP Digital Camera, Canon WP-DC80 Waterproof Case for SD550, ... [Link](http://www.nextag.com/canon-camera/zzulic/biz20/search-html)

Digital Cameras - Reviews, Best Price Comparison and BizRate Store...

This Canon digital camera features a resolution of 5 megapixels ... Offering automatic and manual settings, this Canon digital camera proves perfect for ... [Link](http://www.bizrate.com/digital_cameras/brand-sony/products__at259--390.html)