CWS: A Comparative Web Search System

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

[†]Microsoft Research Asia, Beijing, P.R.China
{jtsun, hjzeng, zhengc}@microsoft.com

[‡]Department of Computer Science,
University of Illinois at Urbana-Champaign
xwang20@cs.uiuc.edu

[§]Department of Computer Science,
Hong Kong University of Science and Technology
dshen@cs.ust.hk

ABSTRACT

In this paper, we define and study a novel search problem: Comparative Web Search (CWS). The task of CWS is to seek relevant and comparative information from the Web to help users conduct comparisons among a set of topics. A system called *CWS* is developed to effectively facilitate Web users' comparison needs. Given a set of queries, which represent the topics that a user wants to compare, the system is characterized by: (1) automatic retrieval and ranking of Web pages by incorporating both their relevance to the queries and the comparative contents they contain; (2) automatic clustering of the comparative contents into semantically meaningful themes; (3) extraction of representative keyphrases to summarize the commonness and differences of the comparative contents in each theme. We developed a novel interface which supports two types of view modes: a pair-view which displays the result in the page level, and a cluster-view which organizes the comparative pages into the themes and displays the extracted phrases to facilitate users' comparison. Experiment results show the *CWS* system is effective and efficient.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-Search Process; H.3.5 [**Information Storage and Retrieval**]: Online Information Services-Web based services

General Terms

Algorithms, Experimentation, Performance

Keywords

Clustering, Comparative Web Search, Keyphrase Extraction, Search Engine

1. INTRODUCTION

Nowadays, search engines have become popular tools for users to seek information from the Web. In general, Web users may have various goals when conducting search. For example, one user may want to find a picture of British Museum, another user may hope to find favorite blogs, and some other users may have the need of comparing two products to guide their purchases. In this paper, we define and study a novel search problem, which we refer to as Comparative Web Search (CWS). CWS is targeted to help users when they wish to make comparisons among a set of topics, e.g., different games, cars, or conferences, etc. Its task is to retrieve relevant and comparative information from the Web so as to facilitate Web users' comparison needs.

Conducting comparisons on the Web is becoming more and more common recently. For example, the emergence of e-commerce makes online shopping very convenient and it is preferred by Web users. To make good purchases, many shoppers indeed first leverage the Web to find relevant information as guidance before their purchases. They may want to compare the features of different products, the online customers' reviews about the products, the stores selling the products, and so on. Other examples include: comparing two related terminologies to understand their differences; comparing two anti-terrorism wars about their costs, their consequences, and also the opinions of the critics. Apparently, CWS can benefit all the above needs.

There are several approaches available which can help people make comparisons on the Web. For example, some newly emerged Web sites began to provide comparison shopping services. Shopping.com and Froogle (http://froogle.google.com) have integrated product comparison services to provide comparative information such as price and customer reviews. However, most of these Web sites are specialized in a certain domain (e.g., products) and can only help fulfill limited comparison tasks for a certain group of users. What's more, their services are based on the structured information provided by the database. Another method is to use traditional search engines for comparative search tasks. Unfortunately, this is not effective since Web users have to manipulate several search windows for a comparative view. To make comparisons with respect to different aspects, users have to frequently refine the queries appropriately or navigate through the result pages. This obviously is tedious for the users. Thus it is much desired to maintain a general platform on which users can easily retrieve and compare every kind of information they need.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

In this paper, we propose a comparative Web search system, *CWS*, which can help users to find comparative information easily. The *CWS* system is different from traditional search engines conceptually. In a traditional search scenario, a Web user submits a query describing his/her information need and a search engine returns a list of presumably relevant pages. In contrast, the objective of our *CWS* system is to facilitate Web users' comparison needs. It allows a user to submit a group of comparative queries with each of them describing a concept the user wants to compare. Our system retrieves the relevant information from the Web, aligns the comparative contents, and ranks them by combining both their relevance to the issued queries and the amount of comparative information they share. Moreover, to help the users digest the comparative contents, we cluster them into different themes and extract representative keyphrases to summarize each theme. At the user end, we implement a novel interface which supports two types of view modes: a pair-view which displays the result in the page level, and a cluster-view which organizes the comparative pages into the themes and displays the extracted phrases to facilitate users' comparison.

In summary, the *CWS* system is characterized by: (1) automatic retrieval and ranking of Web pages based on both their relevance to queries and the comparative contents they contain; (2) automatic clustering the comparative contents into semantically meaningful themes; (3) extraction of representative keyphrases to summarize the commonness and differences of the comparative contents in each theme.

The remainder of this paper is organized as follows. Section 2 provides related works. Section 3 gives a brief introduction to our *CWS* system and we describe our algorithms in Section 4. Section 5 presents the experimental results and Section 6 offers some concluding remarks and directions for future research.

2. RELATED WORK

There were few works on comparative Web search. The most related ones are those focusing on comparing specific Web sites or data collections. Liu et al. [10, 11] compares two Web sites, e.g., the sites of two competitive companies. Given two Web sites, all their pages are merged and partitioned into hierarchical clusters. The pages are then displayed in a tree form and visualization techniques are adopted to emphasize the differences between the two sites. In [12], the authors developed a comparative browser for comparing pages of two Web sites. Their system concurrently presents multiple Web pages thus enabling users to view them at the same time. After a user selects a page from one site, the system retrieves similar contents from the other site. Our system is different from the above works since our purpose is to conduct Web search given a set of comparative queries, instead of making comparisons between two Web sites. Recently, Zang and Zhai et al. define a novel comparative text mining (CTM) problem [21, 18]. Though related, CTM is different from comparative Web search: comparative text mining is conducted on a set of comparative text collections to discover latent common themes across all collections as well as the themes specific to each collection. Tao and Zhai [16] conducted mining on the comparable bilingual text corpus to align a word from one language to a word in another language based on their statistical informtion. In contrast, the task of CWS is query-dependent and the objective is to retrieve comparative information from the Web. Another related work is opinion mining [7, 9]. It is to extract customers' opinions on product features based on a collection of customer reviews. Then both customers and manufactures can make comparisons between products. The authors use several natural language processing techniques and data mining approaches to help identify product features and sentiments of customer opinions. Their methods can not be easily used in CWS because they are domain-dependent. Moreover, the data used in opinion mining is usually well organized and less noisy. All the above works are based on offline mining while CWS focuses on online search.

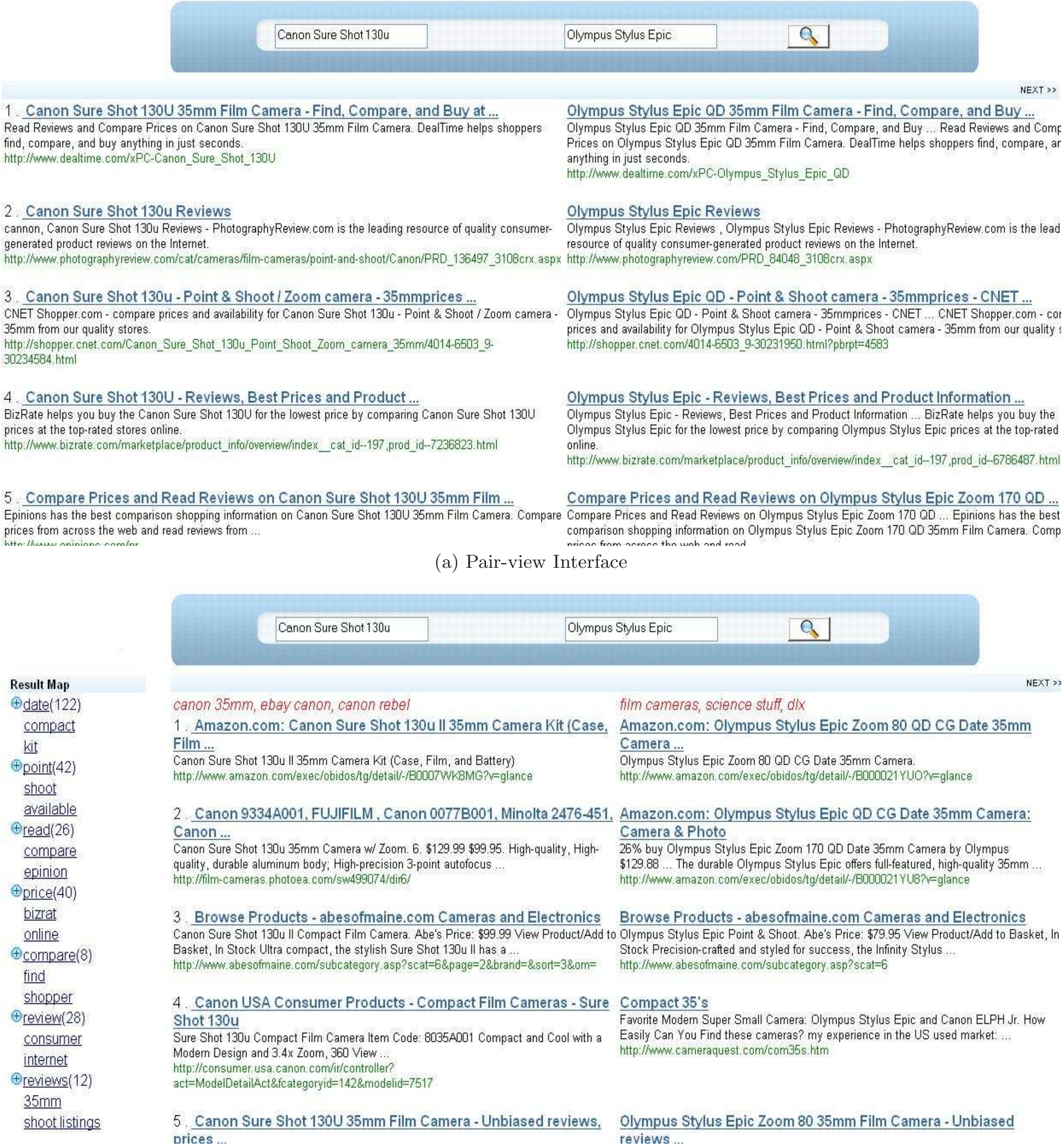
In this paper, we developed a comparative search system named *CWS*. Our system can automatically retrieve Web pages containing comparative information and align comparative page pairs. As far as we know, the available search systems have no such kind of functionalities. Another advantage of our system lies in that it is able to organize the comparative Web pages into clusters and extract keyphrases from them to summarize the common contents of a cluster, as well as the differences between the concepts compared. There are some recent researches on search result clustering [19, 1, 8]. Different from them, our objective is to cluster comparative page pairs in order to facilitate Web users' comparison purpose. In this paper, we adopted a probabilistic clustering algorithm proposed in [21]. The advantage of this approach is that it provides a method to rank the topic themes of all clusters and can produce representative terms for each cluster.

There are also some works on automatic keyphrase extraction from documents [17, 20]. In [17], the authors developed a system named KEA, which uses Naive Bayes algorithm to extract keyphrases. In [20], the authors proposed a simultaneous method for keyphrase extraction and text summarization by modeling text documents as bipartite graphs. In [6], the authors discussed the extraction of important phrases from a text stream (e.g., news) and use it as a query to search relevant pages from the Web. In this paper, we use a keyphrase extraction system, called KEX, developed in our group to extract keyphrases [3]. Furthermore, we also propose an entropy based method to select keyphrases which are unique to the concepts compared by a Web user.

3. SYSTEM OVERVIEW

In this section, we give an overview of our *CWS* system. Figure 1 illustrates the flowchart of our system. For simplicity, our system allows users to give two comparative queries q_1 and q_2 as input. Both queries are submitted to a search engine to get the ranked list of pages from the Web. Then, we re-organize these two lists to get the comparative page pairs and rank them. This is the pair-view output. To help the users to digest the information, we also adopted one clustering algorithm to group the similar pairs together. The keyphrases are extracted from the clusters to highlight the contents of the clusters. This gives the cluster-view output.

Figure 2 gives an example of the *CWS* system interface. The pair-veiw is illustrated in Figure 2(a) and the cluster-view is given in Figure 2(b). In both modes, two text boxes are provided to input the comparative queries. In the pair-view mode, after queries are submitted, two lists of Web pages are generated by the system and are displayed in two columns. The left list of pages correspond to the query con-



Canon Sure Shot 130u

Olympus Stylus Epic

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.dealtime.com/xPC-Canon_Sure_Shot_130U

2 . Canon Sure Shot 130u Reviews
cannon, Canon Sure Shot 130u Reviews - PhotographyReview.com is the leading resource of quality consumer-generated product reviews on the Internet.
http://www.photographyreview.com/cat/cameras/film-cameras/point-and-shoot/Canon/PRD_136497_3108crx.aspx

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_Shot_130u_Point_Shoot_Zoom_camera_35mm/4014-6503_9-30234584.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.
http://www.bizrate.com/marketplace/product_info/overview/index__cat_id--197,prod_id--7236823.html

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 ...
http://www.epinions.com/pr

Olympus Stylus Epic QD 35mm Film Camera - Find, Compare, and Buy ...
Olympus Stylus Epic QD 35mm Film Camera - Find, Compare, and Buy ... Read Reviews and Comp Prices on Olympus Stylus Epic QD 35mm Film Camera. DealTime helps shoppers find, compare, ar anything in just seconds.
http://www.dealtime.com/xPC-Olympus_Stylus_Epic_QD

Olympus Stylus Epic Reviews
Olympus Stylus Epic Reviews , Olympus Stylus Epic Reviews - PhotographyReview.com is the lead resource of quality consumer-generated product reviews on the Internet.
http://www.photographyreview.com/PRD_84048_3108crx.aspx

Olympus Stylus Epic QD - Point & Shoot camera - 35mmprices - CNET ...
Olympus Stylus Epic QD - Point & Shoot camera - 35mmprices - CNET ... CNET Shopper.com - cor prices and availability for Olympus Stylus Epic QD - Point & Shoot camera - 35mm from our quality s
http://shopper.cnet.com/4014-6503_9-30231950.html?pbrpt=4583

Olympus Stylus Epic - Reviews, Best Prices and Product Information ...
Olympus Stylus Epic - Reviews, Best Prices and Product Information ... BizRate helps you buy the Olympus Stylus Epic for the lowest price by comparing Olympus Stylus Epic prices at the top-rated online.
http://www.bizrate.com/marketplace/product_info/overview/index__cat_id--197,prod_id--6786487.html

Compare Prices and Read Reviews on Olympus Stylus Epic Zoom 170 QD ...
Compare Prices and Read Reviews on Olympus Stylus Epic Zoom 170 QD ... Epinions has the best comparison shopping information on Olympus Stylus Epic Zoom 170 QD 35mm Film Camera. Comp prices from across the web and read

(a) Pair-view Interface

Canon Sure Shot 130u

Olympus Stylus Epic

Result Map

- date(122)
  - compact
  - kit
- point(42)
  - shoot
  - available
- read(26)
  - compare
  - epinion
- price(40)
  - bizrat
  - online
- compare(8)
  - find
  - shopper
- review(28)
  - consumer
  - internet
- reviews(12)
  - 35mm
  - shoot listings

canon 35mm, ebay canon, canon rebel

1 . Amazon.com: Canon Sure Shot 130u II 35mm Camera Kit (Case, Film ...
Canon Sure Shot 130u II 35mm Camera Kit (Case, Film, and Battery)
http://www.amazon.com/exec/obidos/tg/detail/-/B0007WK8MG?v=glance

2 . Canon 9334A001, FUJIFILM , Canon 0077B001, Minolta 2476-451, Canon ...
Canon Sure Shot 130u 35mm Camera w/ Zoom. 6. $129.99 $99.95. High-quality, High-quality, durable aluminum body; High-precision 3-point autofocus ...
http://film-cameras.photoea.com/sw499074/dir6/

3 . Browse Products - abesofmaine.com Cameras and Electronics
Canon Sure Shot 130u II Compact Film Camera. Abe's Price: $99.99 View Product/Add to Basket, In Stock Ultra compact, the stylish Sure Shot 130u II has a ...
http://www.abesofmaine.com/subcategory.asp?scat=6&page=2&brand=&sort=3&om=

4 . Canon USA Consumer Products - Compact Film Cameras - Sure Shot 130u
Sure Shot 130u Compact Film Camera Item Code: 8035A001 Compact and Cool with a Modern Design and 3.4x Zoom, 360 View ...
http://consumer.usa.canon.com/ir/controller?act=ModelDetailAct&fcategoryid=142&modelid=7517

5 . Canon Sure Shot 130U 35mm Film Camera - Unbiased reviews, prices ...

film cameras, science stuff, dlx

Amazon.com: Olympus Stylus Epic Zoom 80 QD CG Date 35mm Camera ...
Olympus Stylus Epic Zoom 80 QD CG Date 35mm Camera.
http://www.amazon.com/exec/obidos/tg/detail/-/B000021YUO?v=glance

Amazon.com: Olympus Stylus Epic QD CG Date 35mm Camera: Camera & Photo
26% buy Olympus Stylus Epic Zoom 170 QD Date 35mm Camera by Olympus $129.88 ... The durable Olympus Stylus Epic offers full-featured, high-quality 35mm ...
http://www.amazon.com/exec/obidos/tg/detail/-/B000021YU8?v=glance

Browse Products - abesofmaine.com Cameras and Electronics
Olympus Stylus Epic Point & Shoot. Abe's Price: $79.95 View Product/Add to Basket, In Stock Precision-crafted and styled for success, the Infinity Stylus ...
http://www.abesofmaine.com/subcategory.asp?scat=6

Compact 35's
Favorite Modern Super Small Camera: Olympus Stylus Epic and Canon ELPH Jr. How Easily Can You Find these cameras? my experience in the US used market: ...
http://www.cameraquest.com/com35s.htm

Olympus Stylus Epic Zoom 80 35mm Film Camera - Unbiased reviews ...

(b) Cluster-view Interface

Figure 2: *CWS* System Interface

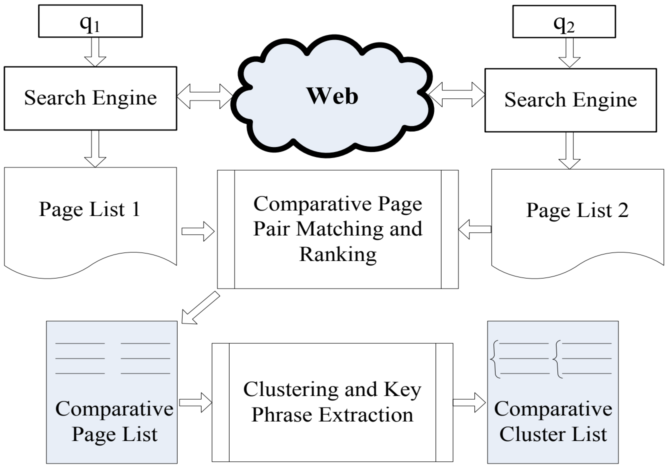


Figure 1: The Flowchart of *CWS* System.

tained in the left textbox, while the right list corresponds to the right query. For each result page, the information including title, URL, and snippet is displayed. There are two differences between the pair-view result and that of traditional search engines. (1) The left page and its corresponding right one share comparative information and they two form a page pair. That is, both pages discuss common topics related to the two input queries. (2) The page pairs are ranked based on their relevance to the queries and the amount of comparative information they contain. In the cluster-view mode, result pages are organized into flat clusters. Each of them contains pages of similar topics. The keyphrases reflecting the common contents of each cluster are extracted and displayed on the left. If a user clicks on these phrases, all the pages of the corresponding cluster will be displayed on the right using the format similar to the pair-view mode. For each of the two page lists in one cluster, the keyphrases unique to this list are also extracted and displayed on the top.

4. ALGORITHMS

Our CWS system is based on an existent search engine, denoted by SE. Given two queries, SE will return two lists of pages ranked by their relevance to the two input queries respectively. We then re-organize the search result pages to facilitate Web users' comparison needs.

4.1 Ranking Comparative Page Pairs

In order to return comparative information for the input queries q_1 and q_2, our first approach is to automatically rerank the search results returned by SE. Assume SR_1 and SR_2 represent the result pages corresponding to queries q_1 and q_2 respectively. In a traditional search, these result pages are ranked by their relevance to the query. In contrast, our purpose is to re-rank SR_1 and SR_2 to display the comparative page pairs. Assume p_1 and p_2 are two pages from SR_1 and SR_2 respectively. If $\langle p_1, p_2\rangle$ is a good comparative pair, p_1 and p_2 should contain information about q_1 and q_2 respectively and both pages should discuss some common aspects about both queries. Our assumption is: if $\langle p_1, p_2\rangle$ is a comparative page pair, they should satisfy:

(1) p_1 is relevant to q_1;
(2) p_2 is relevant to q_2;
(3) If q_1 and q_2 are removed from p_1 and p_2 respectively, the remaining contents of p_1 and p_2 are similar.

We use R to denote the relevance of a query to a page, and S to denote the similarity between two text segments. The function below is used to estimate the likeliness that two pages form a comparative pair with regard to the input queries:

$$\begin{aligned} f_{q_1,q_2}(p_1,p_2) &= \alpha \cdot R(p_1,q_1) \\ &+ \beta \cdot R(p_2,q_2) \\ &+ \lambda \cdot T_{q_1,q_2}(p_1,p_2) \end{aligned} \qquad (1)$$

$$\begin{aligned} T_{q_1,q_2}(p_1,p_2) &= \theta \cdot S(url_1,url_2) \\ &+ (1-\theta) \cdot S(p_1 \backslash q_1, p_2 \backslash q_2) \\ &\forall p_1 \in SR_1, p_2 \in SR_2 \end{aligned} \qquad (2)$$

In Equation (1), $T_{q_1,q_2}(p_1,p_2)$ measures the amount of comparative information of p_1 and p_2 associated with q_1 and q_2. The function f considers the relevance between pages and their corresponding queries, as well as the comparative information contained in the two pages. Parameters α and β are set to be equal in order to guarantee the relevance measures corresponding with the two queries are treated equally. λ is a tradeoff parameter, balancing the relevance measure and the comparison measure. When λ is zero, the above equation is only a linear combination of relevance information. In Equation (2), the comparative information of p_1 and p_2 is computed based on their contents and URLs, with θ balancing the two kinds of information. $p_1 \backslash q_1$ and $p_2 \backslash q_2$ denote the remaining text contents of page p_1 and p_2 after removing terms contained in their snippet texts respectively. $S(url_1, url_2)$ denotes the similarity between the URL strings of p_1 and p_2.

The computation of f is straightforward. In traditional search, R is used to rank Web pages. Usually two factors are considered: the first is the importance of a page, which is usually computed based on the links among Web pages (e.g. PageRank [13]); the second is the similarity between a query and a page, which can be computed by traditional information retrieval models, such as probabilistic model, vector space model, etc, [2]. These models can also be used for the computation of S.

It is quite common for a page editor to put some comparative contents about q_1 and q_2 in one single page. Such kinds of pages will be in both SR_1 and SR_2. In this paper, we regard these kinds of pages themselves as comparative pages. The ranking of these pages can also be handled by our approach. In this case, $T_{q_1,q_2}(p_1,p_2)$ is maximal because the same contenst are left if q_1 and q_2 are removed from the original pages and both pages share the same URL. Thus only $\alpha \cdot R(p_1,q_1) + \beta \cdot R(p_2,q_2)$ is needed for ranking purpose.

Our purpose is to identify the comparative page pairs from the pages of SR_1 and SR_2. Those pages form a bipartite graph, where the edge weight is computed by f. Although traditional maximum matching algorithms can also be used to for pair matching [14], they are not suitable for the comparative search task for two reasons: 1) The maximum matching algorithms are not efficient, while CWS is an online application. 2) When Web users make comparisons in a search scenario, they are usually interested in the top results rather than the whole list. Thus it is unnecessary to find a group of page pairs based on maximizing an objective function. In this paper, we proposed a greedy algorithm to rank the comparative page pairs, as discussed below.

All page pairs $E = \{\langle p_1, p_2\rangle | p_1 \in SR_1, p_2 \in SR_2\}$ are first ranked in decreasing order according to $f_{q_1,q_2}(p_1, p_2)$. The pair with the highest score will be selected as a comparative pair and both pages of this pair are inserted in set P. All the remaining page pairs will be filtered and those containing pages in P are removed from E. Then the second comparative pair is selected from the updated set E. This process iterates until E is empty. With this strategy, we can remove those pairs containing duplicate pages and rank all the comparative page pairs according to f.

4.2 Clustering Comparative Page Pairs

In Section 4.1, we did not consider the redundancy among the comparative page pairs. That is, there may exist several page pairs describing the similar aspects of the two input queries. For example, all the comparative pairs ranked at top may compare the prices of two products, thus users have to navigate down through the pair list to find comparative contents about other aspects. In order to address this problem, we propose a comparative page clustering approach to improve the comparison results. At the user end, we present comparative page clusters instead of page pairs. Each cluster consists of pages describing similar aspect(s) of the comparative contents. Pages in a cluster c are divided into two parts: c_1 and c_2, which contains contents specific to q_1 and q_2 respectively.

We cluster the comparative page pairs produced in Section 4.1 to generate the comparative clusters. Each page pair $\langle p_1, p_2\rangle$ is treated as a whole consisting of all the snippets associated with p_1 and p_2. Then all the page pairs are clustered by a probabilistic clustering algorithm. For each cluster, its page pairs are displayed side by side for comparison purpose.

The clustering algorithm is based on the simple mixture generative model [21]. In the mixture generative model, each document is generated by a mixture of several multinomial word distributions. These word distributions correspond with the latent themes among all documents and can be estimated by the Expectation-Maximization (EM) algorithm [4]. At the same time, the EM algorithm can also give us the mixing weights of each document to the themes (i.e., word distributions). The document clusters are then formed by assigning each document to the most salient theme to which it has the largest weight.

Formally, assume there are k hidden themes in a given document collection C: $\theta_1, \cdots, \theta_k$, and one background model θ_B which has high probability to generate the common English words such as "the" and "a". A document d is regarded as a sample of the following mixture model:

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

where w is a word, $\pi_{d,j}$ is the document mixing weight associated with the j-th theme and $\sum_{j=1}^{k} \pi_{d,j} = 1$, and λ_B is the mixing weight for the background model. To estimate the parameters $\Omega = \{\pi_{d,j}, \theta_j | d \in C, j = 1, \cdots, k\}$, the log-likelihood of the collection is defined:

$$logP(C|\Omega) = \sum_{d \in C, w \in V} c(w, d) \cdot logP(w|\theta_d)$$

where V is the vocabulary, $c(w, d)$ is the count of word w in document d. The purpose is to find good parameters to maximize the log-likelihood and it can be achieved by a standard EM algorithm. More details about the EM algorithm can be found in [21]. After the document d's mixing weights to each theme model are achieved, d can be assigned to the cluster by

$$\hat{j} = argmax_j \{\pi_{d,j} | j = 1, \cdots, k\}$$

From the estimated word distribution $P(w|\theta_j)$, the most important words for the j-th theme can be selected by incorporating their probabilities in θ_j. These words are representative of the j-th theme and will be displayed in our CWS system for the j-th cluster. In our system, the clusters are ranked based on their salience scores $\frac{1}{|C|} \sum_{d \in C} \pi_{d,j}$.

4.3 Extracting Keyphrases for Comparative Clusters

After the page pairs are clustered, we extract keyphrases from each cluster in order to facilitate users' comparisons. As each cluster consists of pages corresponding with two queries, we extract the phrases reflecting the common theme of all these pages in one cluster, as well as those specific to each query. As discussed in Section 4.2, the important words estimated by the clustering algorithm will be used as common keyphrases for each cluster. In this section, we first describe our approach to extracting keyphrases for each page. Then we discuss our entropy based method to select keyphrases specific to each query from the phrases generated in the previous step.

4.3.1 Keyphrase Extraction Algorithm

We use a phrase extraction package, KEX, implemented in our group to extract keyphrases for each result page [3]. KEX is based on a supervised approach. The training examples of our package are created by three human annotators who manually extract keyphrases from a collection of Web pages. For each candidate phrase in a Web page, a 4-dimensional feature vector $\langle x_1, x_2, x_3, x_4\rangle$ is constructed. These phrases are used to train a linear regression model:

$$y = b_0 + \sum_{i=1}^{4} b_i x_i \tag{3}$$

If a phrase is keyphrase, $y = 1$; otherwise, $y = 0$. The phrase features include:

(1) *PF*: phrase frequency. This feature is calculated in the traditional meaning of term frequency (*TF*). Intuitively, frequent phrases are more likely to be better candidates of keyphrases.

(2) *ATF*, average frequency of all terms in the phrase. Sometimes, a keyphrase may have low PF but contain keyterms with high TF. The *ATF* feature can be used to discover this kind of keyphrases.

(3) *AIDF*, average inverse document frequency (*IDF*) of all terms contained in the phrase. Intuitively, if a phrase contains many terms with low IDF, it is less informative.

(4) *OKA*, modified Okapi weighting score. Okapi is a

highly effective document weighting model in information retrieval [5]. The formula is:

$$\sum_{w\in q\cap d} ln\frac{N-df(w)+0.5}{df(w)+0.5} \times \frac{(k_1+1)\times c(w,d)}{k_1((1-b)+b\frac{|d|}{avdl})+c(w,d)} \times \frac{(k_3+1)\times c(w,d)}{k_3+c(w,d)} \qquad (4)$$

In our system, we adopt this parameter setting: $k_1 = 1.2$, $b = 0.25$, $k_3 = 1000$ and $avdl = 100$. We use *log(OKA)* score as a feature.

After the feature vectors are constructed for all the candidate phrases, we train a linear regression model as described in Equation (3), where x_1=*PF*, x_2=*ATF*, x_3=*AIDF* and x_4=*log(OKA)*. Then we apply this model on every page in each cluster c to rank all candidate phrases. Those ranked at top are selected as keyphrases. In our system, all the candidate phrases are extracted from the title and snippet text returned by SE. We do not use the HTML contents to guarantee the efficiency of our CWS system, as downloading these pages and parsing them are quite time-consuming.

4.3.2 Keyphrase Selection for Clusters

As the query specific keyphrases summarize the contents contained in sub-clusters c_1 and c_2 respectively. We propose to use the entropy measure to help select them.

$$Ent(w) = -\sum_{i=1,2} p_i \log p_i$$

where p_i(i=1,2) measures the probability that phrase w occurs in sub-cluster $c_i(i = 1, 2)$. For each sub-cluster, all the keyphrases contained in it are ranked by $Ent(w)$ and those with low entropies are regarded as query specific phrases. Intuitively, if a phrase frequently occurs in one sub-cluster and seldom occurs in the other, it has low entropy value and will be regarded as a keyphrase specific to the current sub-cluster.

5. EXPERIMENTS

In this section, we investigate whether our CWS system can help to satisfy Web users' comparison needs. Both the pair-view and the cluster-view modes are used for experiments. Twenty pairs of comparative queries listed in Table 1 are used. We intentionally select the query pairs broadly which reflect different comparison needs: cameras, companies, diseases, and humans, etc.

For evaluation purpose, three human subjects are requested to annotate all the 20 query pairs. For each query pair, we submit them to MSN search engine and retrieve at most the top 50 pages for each query. Each subject is asked to navigate through the snippet texts of the 100 pages and manually match the comparative page pairs. If two pages satisfy the below 3 conditions, they will be labeled by a subject as a comparative page pair:

1) The first page is relevant with the first query.

2) The second page is relevant with the second query.

3) The contents of both pages can help users make comparisons. The labeling results of all three subjects are used to evaluate our CWS system.

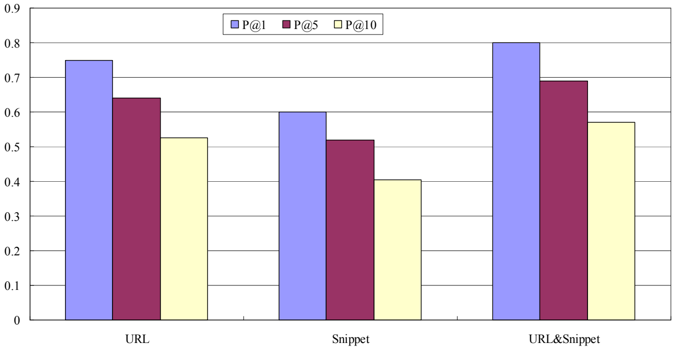


Figure 3: Precision measures of comparative page pair results

5.1 Results of the Comparative Page Pair Ranking Approach

In this experiment, we evaluate the effectiveness of the comparative page pairs returned by CWS in the pair-view mode. As discussed in Section 4.1, we need to compute R and T to rank the page pairs by f. In this experiment, as we use a search engine to retrieve Web pages, the search engine does not return the relevance score between a query and a page. We have only the rank order of the result pages. A straightforward approach to estimate the relevance between a query q and a page p is: $R(q,p) = \frac{1}{r}$, r is the rank of the page in the corresponding search results returned by SE. The cosine similarity is used to compute the function T in Equation (2) [2].

5.1.1 Results Measured by Precision

Based on the annotated results, we can calculate the precision measures $P@N$ of the comparative pair ranking results. For each page pair, $P@N$ is defined as the number of comparative page pairs in the top N pair results divided by N. In our experiment, N take values 1, 5 and 10.

In Equation (1), the parameters α, β and λ may influence both the construction of page pairs and their ranks in the result. In order to give equal weights to both queries, we set $\alpha = \beta$ and require $\alpha + \beta + \lambda = 1$. Since both R and T functions take values from 0 to 1, we vary all possible values of the parameters α, β, λ and θ and report the best result achieved by our system. In our experiments, the values of the above parameters are varied from 0 to 1 with step 0.1.

As given in Equation (2), the comparative information of two pages is calculated using their snippet texts and URLs. In order to compare their effectiveness, we also report the results when only one kind of information is used. As given in Figure 3, "URL" corresponds with $\theta = 1$ in Equation (2), "Snippet" corresponds with $\theta = 0$ and "URL&Snippet" denotes both kinds of information are used. In all our experiments, the snippet text of a page is the combined strings of its title and the snippet returned by the search engine. In the three cases, all possible parameters are varied and the best pair ranking result is reported in Figure 3. For each setting, the evaluation results of P@1, P@5 and P@10 are all given. All the precision measures are averaged over the annotation results of the three subjects.

From Figure 3, we can find both the URL and the snippet information are useful when calculating the comparative information of two Web pages. When only one kind of in-

Table 1: 20 Pairs of Comparative Queries

	q_1	q_2		q_1	q_2
1	xbox	playstation	11	sars	bird flu
2	Sony dv	Samsung dv	12	McDonalds	KFC
3	Canon sure shot 130u	Olympus stylus epic	13	Nike	Adidas
4	lancome	clinique	14	Iraq war	Afghanistan war
5	Ford Escape	Jeep Liberty	15	virtual earth	Google map
6	PSP	GBA	16	Chengxiang Zhai	Jiawei Han
7	Maradona	Pele	17	Sony Camera	Canon camera
8	Nokia cell phone	Motorola cell phone	18	windows	linux
9	MIT	CMU	19	MSN messenger	Google talk
10	Microsoft	Google	20	Bush	Clinton

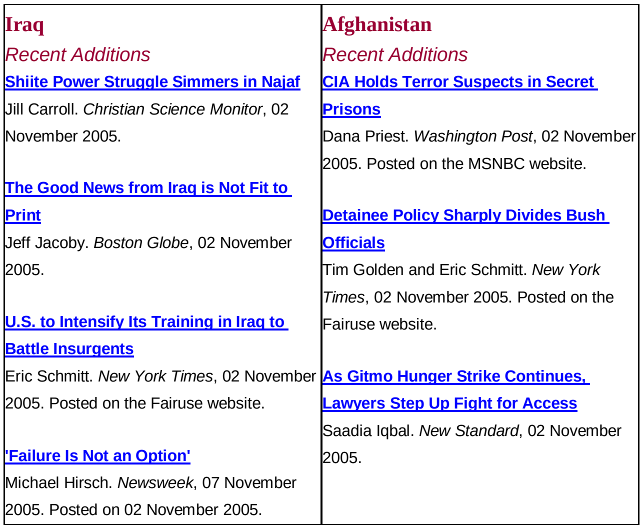
formation is used, "URL" is better than "Snippet". The combination of them leads to better comparative ranking results. The conclusions are consistent when the results are evaluated by P@1, P@5 and P@10 respectively. The best P@10 (in "URL&Snippet" setting) precision is 0.57, which indicates 57% page pairs in the top 10 results returned by our *CWS* system are meaningful comparative page pairs.

5.1.2 Case Studies

In Section 5.1.1, the effectiveness of comparative page pairs are evaluated using precision measure. Here, we also study two cases in order to give intuitive results of our *CWS* system.

In Table 2, we give the results of two query pairs. The first pair contains two product queries: 'Canon Sure Shot 130u' and 'Olympus Stylus Epic'. The second consists of query 'Afghanistan War' and 'Iraq War'. The titles and URLs of each page pair are given side by side but the snippets are omitted for the limit of space.

The two product queries refer to two types of cameras manufactured by *Cannon* and *Sony*, respectively. Web users may submit these two queries in order to make comparisons between the two cameras. From the annotation results, we find that all the three subjects annotate the 10 results as comparative page pairs. As listed in Table 2, for the first 9 page pairs, both pages of each pair come from a same website. Take the first pair as an example: *DealTime* (http://www.dealtime.com/) is an online shopping Web site and the two pages in this pair come from this website. Both pages contain the price information of several shops selling the corresponding cameras. The two pages are automatically discovered by our system and form a comparative pair. As for the second page pair, *PhotographReview* (http://www.photographyreview.com/) is a site providing information like digital camera and photo equipment reviews. The pages returned by our system are exactly the two containing the customer reviews about the two cameras queried by the user. The next 7 pages are also comparative page pairs of other Web sites. That is, our *CWS* system can integrate the comparative pages of various Web sites together and present them to end users, which will greatly facilitate Web users' comparison needs. As for the 10th pair returned by our system, the two pages come from *Shopping.com* and *DealTime*, respectively, and are put together to form a comparative page pair. This indicates the pages from different Web sites can also be identified to form a comparative page pair.



Iraq

Recent Additions

Shiite Power Struggle Simmers in Najaf

Jill Carroll. *Christian Science Monitor*, 02 November 2005.

The Good News from Iraq is Not Fit to Print

Jeff Jacoby. *Boston Globe*, 02 November 2005.

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

Eric Schmitt. *New York Times*, 02 November 2005. Posted on the Fairuse website.

'Failure Is Not an Option'

Michael Hirsch. *Newsweek*, 07 November 2005. Posted on 02 November 2005.

Afghanistan

Recent Additions

CIA Holds Terror Suspects in Secret Prisons

Dana Priest. *Washington Post*, 02 November 2005. Posted on the MSNBC website.

Detainee Policy Sharply Divides Bush Officials

Tim Golden and Eric Schmitt. *New York Times*, 02 November 2005. Posted on the Fairuse website.

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

Saadia Iqbal. *New Standard*, 02 November 2005.

Figure 4: A comparative page returned for query pair: 'Afghanistan war' and 'Iraq war'.

Table 2 also gives the results for the query pair: 'Afghanistan war' and 'Iraq war'. Web users may submit the two queries in order to make comparisons between the two recent wars. We can find that the 5th page pair consists of only one page. This page should contain comparative contents relevant with both wars. This is verified after we check this page. It is a war report page which archives articles about the two wars. All the articles are listed side by side, the left corresponding with the Iraq war and the right corresponding with the Afghanistan war. Partial contents of this page are displayed in Figure 4.

5.2 Results of Comparative Page Clustering and Keyphrase Extraction

Traditional document clustering relies on the category information as ground truth for evaluation [15]. However there is no such information for all the pages we clustered. Instead, we evaluate the clustering results by investigating the accuracy of the extracted keyphrases.

The KEX package is used to extract keyphrases for each result page [3]. The linear regression model is trained on a set of 300 Web pages which have been manually annotated by three human subjects. This model can achieve a top 10

Table 2: Results Returned by *CWS* in Pair-view Mode

	q_1=‘**Canon Sure Shot 130u**’, q_2=‘**Olympus Stylus Epic**’	
1.	Canon Sure Shot 130U 35mm Film Camera - Find, Compare, and Buy at ... http://www.dealtime.com/xPC-Canon_Sure_Shot_130U	Olympus Stylus Epic QD 35mm Film Camera - Find, Compare, and Buy ... http://www.dealtime.com/xPC-Olympus_Stylus_Epic_QD
2.	Canon Sure Shot 130u Reviews http://www.photographyreview.com/cat/cameras/film-cameras/point-and-...	Olympus Stylus Epic Reviews http://www.photographyreview.com/PRD_84048_3108crx.aspx
3.	Canon Sure Shot 130u - Point & Shoot / Zoom camera - 35mmprices ... http://shopper.cnet.com/Canon_Sure_Shot_130u_Point_Shoot_Zoom	Olympus Stylus Epic QD - Point & Shoot camera - 35mmprices - CNET ... http://shopper.cnet.com/4014-6503_9-30231950.html?pbrpt=4583
4.	Canon Sure Shot 130U - Reviews, Best Prices and Product ... http://www.bizrate.com/marketplace/product_info/overview/index...	Olympus Stylus Epic - Reviews, Best Prices and Product Information ... http://www.bizrate.com/marketplace/product_info/overview/index__cat_id...
5.	Compare Prices and Read Reviews on Canon Sure Shot 130U 35mm Film ... http://www.epinions.com/pr-Film_Cameras_Canon_Sure_Shot_130u_Ca...	Compare Prices and Read Reviews on Olympus Stylus Epic Zoom 170 QD ... http://www.epinions.com/pr-Film_Cameras_Olympus_Stylus_Epic_Zoom_170...
6.	Canon Sure Shot 130u II 35mm Camera Kit @ Unverse http://www.unverse.com/id-Canon+Sure+Shot+130u+II+35mm+Came...	Olympus Stylus Epic Zoom 170 QD Date 35mm Camera @ Unverse http://www.unverse.com/id-Olympus+Stylus+Epic+Zoom+170+QD+Da...
7.	Compare Prices and Read Reviews on Canon Sure Shot 130U 35mm Film ... http://www.epinions.com/pr-film_cameras_canon_sure_shot_130u_caption_35mm_p...	Compare Prices and Read Reviews on Olympus Stylus Epic DLX 35mm ... http://www.epinions.com/elec_Cameras-Point_And_Shoot_OlympusStyluss-...
8.	Canon Sure Shot 130u - Point & Shoot / Zoom camera - 35mm - SLR ... http://www.mysimon.com/Canon_Sure_Shot_130u_Point_Shoot_Zoom_cam...	Olympus Stylus Epic QD - Point & Shoot camera - 35mm - SLR ... http://www.mysimon.com/Olympus_Stylus_Epic_QD_Point_Shoot_camera_...
9.	Canon Sure Shot 130u 35mm Camera w/ Zoom @ Unverse http://www.unverse.com/id-Canon+Sure+Shot+130u+35mm+Camera+w+Zoom-B00006K154	Olympus Stylus Epic QD CG Date 35mm Camera @ Unverse http://www.unverse.com/id-Olympus+Stylus+Epic+QD+CG+Date+35mm...
10.	Canon Sure Shot 130U 35mm Film Camera - Find, Compare, and Buy at ... http://www.shopping.com/xPC-Canon_Sure_Shot_130U	Olympus Stylus Epic Zoom 170 QD 35mm Film Camera - Find, Compare ... http://www.dealtime.com/xPC-Olympus_Stylus_Epic_Zoom_170_QD
	q_1=‘**Afghanistan War**’, q_2=‘**Iraq War**’	
1.	Afghanistan War . The Columbia Encyclopedia, Sixth Edition. 2001-05 http://www.bartleby.com/65/af/AfghanWar.html	Iran- Iraq War . The Columbia Encyclopedia, Sixth Edition. 2001-05 http://www.bartleby.com/65/ir/IranIraq.html
2.	The Observer — Special reports — War in Afghanistan http://observer.guardian.co.uk/afghanistan/0,1501,573451,00.html	Muslims, Islam, and Iraq http://www.uga.edu/islam/iraq.html
3.	Afghanistan Timeline, 21st Century http://www.mapreport.com/countries/afghanistan.html	Iraq War Timeline http://www.infoplease.com/ipa/A0908792.html
4.	Articles about September 11 2001 attacks on USA and subsquent ... http://people.pwf.cam.ac.uk/nwm20/usa_afghanistan.htm	Iraq War http://webhost.bridgew.edu/jhayesboh/iraq.html
5.	War Report - Iraq War and Afghan Aftermath - compiled by the ... http://www.comw.org/warreport/	
6.	Independent Online Edition > World Politics: http://news.independent.co.uk/world/politics/article313450.ece	Informed Comment http://www.juancole.com/2004/07/preoccupation-with-iraq-slowed-us-uk.html
7.	Government Resources http://library.louisville.edu/government/subjects/war/afgwar/afgwar.html	VAIW :: Veterans Against The Iraq War http://www.vaiw.org/vet/index.php
8.	events 19691979 crises recovery eec world renewal tensions cartoon ... http://www.ena.lu/europe/crisis-recovery/cartoon-murschetz-afghanistan-war.htm	Iraq War Cartoons http://www.cartoonistgroup.com/bysubject/theiraqcartoons.php
9.	Amazon.com: The Lessons of Afghanistan : War Fighting, Intelligence ... http://www.amazon.com/exec/obidos/tg/detail/-/089206417X?v=glance	Amazon.com: The Iraq War : Books http://www.amazon.com/exec/obidos/tg/detail/-/1400041996?v=glance
10.	Afghanistan : War Without End? http://www.pbs.org/newshour/bb/asia/afghanistan/afghan_12-27-85.html	UNCOVERED: The War on Iraq http://www.truthuncovered.com/

precision and recall of 0.303 and 0.297 respectively. This result is not bad, because when we evaluate the annotation result of one subject on those of the other two, the average precision at 10 and the recall at 10 is 0.44 and 0.388 respectively. These values indicate that keyphrase extraction is quite subjective and not an easy task. This conclusion is also drawn in previous research works [17]. In this paper, we do not present the evaluation details of our keyphrase extraction algorithm.

Table 3 presents the phrases extracted for query pair: 'ChengXiang Zhai' and 'Jiawei Han'. Table 4 corresponds the result of query pair: 'Canon Sure Shot 130u' & 'Olympus Stylus Epic'. For each cluster, the top 3 common keyphrases as well as the top 3 keyphrases specific to each query are given. As we extract the query specific keyphrases, those which are sub phrases of the query are omitted as they do not provide additional information.

The result given in Table 3 is very interesting. As both the professors are from University of Illinois at Urbana-Champaign, from the three common phrases we can find that the first cluster corresponds with the pages introducing the two professors. The second cluster corresponds with their research works and the third is about their publications. Most query specific phrases also make sense. For example, in the third and fourth clusters, phrases like "information retrieval" are extracted for the query "ChengXiang Zhai" and phrases such as "data mining" are extracted for the query "Jiawei Han". This exactly reflects the different research interests between Professor ChengXiang Zhai and Professor Jiawei Han.

As for the results of the two camera queries, the results are also interesting. For the first cluster, the words 'date', 'compact' and 'kit' are extracted as common keyphrases. This is because both the cameras are compact. The two terms 'date' and 'kit' also frequently appear in all the result pages corresponding with the two queries. According to the common phrases, we can also find that clusters 3, 6 and 7 contain pages on consumer reviews and cluster 4 is about price comparisons.

5.3 Discussions

Based on the above experiments and case studies, we find our *CWS* system is effective. In the pair-view mode, the percentage of meaningful comparative page pairs in the top 1, 5, 10 results is 80%, 69% and 57% respectively. We can also find the combination of URL and snippet contents is effective in measuring the comparative information of two pages. The case studies also show our comparative page ranking function is able to find those pages which contain comparison information relevant with both input queries.

As Equation (1) indicates, both the comparative and relevance information help decide whether two pages form a meaningful comparative pair. We also did experiments to study which kind of information is more promising. In this experiment, the parameter θ is fixed and α, β and λ are varied. The conclusion is: with the increase of λ, the precison of the pair matching grows steadily. This shows the relevance information between queries and pages has no impact on the pair matching result. The reason is: when the three subjects annotated the 20 queries, they only identified which two pages form a comparative pair. They did not rank the pairs according to their relevance scores with the input queries. When λ is small, even if those comparative page pairs which are very relevant with the input queries can be identified, they do not make extra contribution to the precison evaluation. At the beginning of the labeling process, we also asked the subjects to rank the comparative page pairs. However, we found ranking them is much more difficult than just identifying whether two pages form a comparative pair or not. Thus we need other approaches to evaluate the ranking order of the comparative page pairs.

In the cluster-view mode, our *CWS* system can automatically cluster the comparative information into different themes. The keyphrases are also extracted to summarize the commonness and differences of each theme. The examples given in Section 5.2 show the comparative information produced by *CWS* are helpful for making comparisons. However, it is hard to quantitively evaluate the clustering results as well as the extracted keyphrases.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed and studied a novel search problem, Comparative Web Search. We developed a *CWS* system to help users seek comparative information from the Web. Human evaluations and some case studies show that our system is quite effective to facilitate users' comparative information needs.

In the future, we plan to investigate the following issues:

(1) The evaluation of the comparative Web search system is challenging and labor intensive. In this paper, our evaluation result of the *CWS* system is based on a relatively small query sets. It is interesting to adopt other approaches to evaluate the effectiveness of comparative search system.

(2) The queries input to the *CWS* system represent the topics which the users will compare. How to automatically distinguish comparative query pairs is also an interesting problem.

(3) In this paper, we combine the contents and the ranking information of Web pages to construct comparative page pairs. We also plan to incorporate the link structure information to our system.

(4) Our approaches to the comparative Web search problem are still preliminary and our *CWS* system only provides very basic comparison functionalities. More advanced functions can be added by leveraging other relevant techniques.

In conclusion, the *CWS* system is challenging but very helpful to satisfy users' comparison needs. We expect to conduct more research work on this direction.

7. ACKNOWLEDGMENTS

We thank Dr. ChengXiang Zhai for insightful discussions and Liu Xin for his help on organizing the labeling work and implementing the *CWS* system. We also thank the reviewers for their valuable suggestions on this work.

8. REFERENCES

[1] Vivisimo website. http://vivisimo.com.

[2] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] M. Chen, J.-T. Sun, H.-J. Zeng, and K.-Y. Lam. A practical system of keyphrase extraction for web pages. In *CIKM*, pages 277–278, 2005.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM

Table 3: Keyphrase Extraction Result for Query Pair: q_1=‘ChengXiang Zhai’, q_2=‘Jiawei Han’

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

Table 4: Keyphrase Extraction Result for Query Pair: q_1=‘Canon Sure Shot 130u’, q_2=‘Olympus Stylus Epic’

	Common Keyphrases	q_1 Specific Keyphrases	q_2 Specific Keyphrases
1.	date, compact, kit (122)	canon 35mm, ebay canon, canon rebel	film cameras, science stuff, dlx
2.	point, shoot, available (42)	compare, canon buy, compact	zoom, resnick, rambling
3.	read, compare, epinion (26)	cameras, shot 130u caption, canon 8036a006	dlx, electronic equipment, glorianas court
4.	price, bizrat, online (40)	photo, shot 130u 35mm camera, photo canon	digital, save, day
5.	compare, find, shopper (8)	35mm film, shot 130u 35mm film camera, cameras	camera-mint, camera, compare
6.	review, consumer, internet (28)	film camera, watch, digital video	equipment used, rooks archives, cg
7.	reviews, 35mm, shoot listings (12)	shoot, reviews canon, 35mm compact	excite partner, photograph, outdoor photographer

algorithm. *J. of the Royal Statistical Society, Series B*, 34:1–38, 1977.

[5] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of SIGIR ’04*, pages 49–56, 2004.

[6] M. R. Henzinger, B.-W. Chang, B. Milch, and S. Brin. Query-free news search. In *WWW ’03: Proceedings of the 12th International Conference on World Wide Web*, pages 1–10, 2003.

[7] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of KDD ’04*, pages 168–177, 2004.

[8] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *Proceedings of WWW ’04*, pages 658–665, 2004.

[9] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of WWW ’05*, pages 342–351, 2005.

[10] B. Liu, Y. Ma, and P. S. Yu. Discovering unexpected information from your competitors’ web sites. In *Proceedings of KDD ’01*, pages 144–153, 2001.

[11] B. Liu, K. Zhao, and L. Yi. Visualizing web site comparisons. In *Proceedings of WWW ’02*, pages 693–703, 2002.

[12] A. Nadamoto and K. Tanaka. A comparative web browser (CWB) for browsing and comparing web pages. In *Proceedings of WWW ’03*, pages 727–735, 2003.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[14] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity.* Prentice Hall, Englewood Cliffs, N.J., 1982.

[15] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *TextMining Workshop, KDD*, 2000.

[16] T. Tao and C. Zhai. Mining comparable bilingual text corpora for cross-language information integration. In *Proceeding of KDD ’05*, pages 691–696, 2005.

[17] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. KEA: Practical automatic keyphrase extraction. In *ACM DL*, pages 254–255, 1999.

[18] P. Zang. CTMS: A comparative text mining system. Master thesis, University of Illinois at Urbana-Champaign, Computer Science Department, 2004.

[19] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *Proceedings of SIGIR ’04*, pages 210–217, 2004.

[20] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of SIGIR ’02*, pages 113–120, 2002.

[21] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of KDD ’04*, pages 743–748, 2004.