User Interfaces and Mobile Web

Maryam Kamvar, Melanie Kellar, Rajan Patel and Ya Xu. Computers and iPhones and Mobile Phones, oh my! A logs-based comparison of search users on different devices.

Abstract: We present a logs-based comparison of search patterns across three platforms: computers, iPhones and conventional mobile phones. Our goal is to understand how mobile search users differ from computer-based search users, and we focus heavily on the distribution and variability of tasks that users perform from each platform. The results suggest that search usage is much more focused for the average mobile user than for the average computer-based user. However, search behavior on high-end phones resembles computer-based search behavior more so than mobile search behavior. A wide variety of implications follow from these findings. First, there is no single search interface which is suitable for all mobile phones. We suggest that for the higherend phones, a close integration with the standard computer-based interface (in terms of personalization and available feature set) would be beneficial for the user, since these phones seem to be treated as an extension of the users' computer. For all other phones, there is a huge opportunity for personalizing the search experience for the user's "mobile needs", as these users are likely to repeatedly

Yu Zheng, Lizhu Zhang, Xing Xie and Wei-Ying Ma. Mining Interesting Locations and Travel Sequences from GPS Trajectories for Mobile Users

Abstract: The increasing availability of GPS-enabled devices is changing the way people interact with the Web, and brings us a large amount of GPS trajectories representing people’s location histories. In this paper, based on multiple users’ GPS trajectories, we aim to mine interesting locations and classical travel sequences in a given geospatial region. Here, interesting locations mean the culturally important places, such as Tiananmen Square in Beijing, and frequented public areas, like shopping malls and restaurants, etc. Such information can help users understand surrounding locations, and would enable travel recommendation. In this work, we first model multiple individuals’ location histories with a tree-based hierarchical graph (TBHG). Second, based on the TBHG, we propose a HITS (Hypertext Induced Topic Search)-based inference model, which regards an individual’s access on a location as a directed link from the user to that location. This model infers the interest of a location by taking into account the following three factors. 1) The interest of a location depends on not only the number of users visiting this location but also these users’ travel experiences. 2) Users’ travel experiences and location interests have a mutual reinforcement relationship. 3) The interest of a location and the travel experience of a user are relative values and are region-related. Third, we mine the classical travel sequences among locations considering the interests of these locations and users’ travel experiences. We evaluated our system using a large GPS dataset collected by 107 users over a period of one year in the real world. As a result, our HITS-based inference model outperformed baseline approaches like rank-by-count and rank-by-frequency. Meanwhile, when considering the users’ travel experiences and location interests, we achieved a better performance beyond baselines, such as rank-by-count and rank-by-interest, etc.

Yuki Arase, Xing Xie, Manni Duan, Takahiro Hara and Shojiro Nishio. A Game Based Approach to Assign Geographical Relevance to Web Images

Abstract: Geographical context is very important for images. Millions of images on the Web have been already assigned latitude and longitude information. Due to the rapid proliferation of such images with geographical context, it is still difficult to effectively search and browse them, since we do not have ways to decide their relevance. In this paper, we focus on the geographical relevance of images, which is defined as to what extent the main objects in an image match landmarks at the location where the image was taken. Recently, researchers have proposed to use game based approaches to label large scale data such as Web images. However, there is no in-depth study on the quality of collected game logs and how the logs can improve existing applications. To answer these questions, we design and implement a Web-based and multi-player game to collect human knowledge while people are enjoying the game. Then we thoroughly analyze the game logs obtained during a three week study with 147 participants and propose methods to determine the image geographical relevance. In addition, we conduct an experiment to compare our methods with a commercial search engine. Experimental results show that our methods dramatically improve image search relevance. Furthermore, we show that we can derive geographically relevant objects and their salient portion in images, which is valuable for a number of applications such as image location recognition.

Joshua Hailpern, Loretta Guarino Reid, richar Boardman and Srinivas Annam. WEB 2.0: BLIND TO AN ACCESSIBLE NEW WORLD

Abstract: With the advent of Web 2.0 technologies, Websites have evolved from static pages to dynamic, interactive Web-based applications with the ability to replicate common desktop functionality. However, for blind and visually impaired individuals who rely upon screen readers, Web 2.0 applications force them to adapt to an inaccessible use model. Many technologies, including WAI-ARIA, AJAX, and improved screen reader support, are rapidly coming together. However, simply combining them does not solve the problems of screen reader users. The main contributions of this paper are two models of interaction for screen reader users, for both traditional Websites and Web 2.0 applications. Further contributions are a discussion of accessibility difficulties screen reader users encounter when interacting with Web 2.0 applications, a user workflow design model for improving Web 2.0 accessibility, and a set of design requirements for developers to ease the user's burden and increase accessibility. These models, accessibility difficulties, and design implications are based directly on responses and lessons learned from usability research focusing on Web 2.0 usage and screen reader users. Without the conscious effort of Web engineers and designers, most blind and visually impaired users will shy away from using new Web 2.0 technology in favor of desktop based applications.

Cameron Braganza, Kim Marriott, Peter Moulder, Michael Wybrow and Tim Dwyer. Scrolling Behaviour with Single- and Multi-column Layout

Abstract: The standard layout model used by web browsers is to lay text out in a vertical scroll using a single column. The horizontal-scroll layout model - in which text is laid out in columns whose height is set to that of the browser and the viewer scrolls horizontally - seems well-suited to multi-column layout on electronic devices. We describe a study that examines how people read and in particular the strategies they use for scrolling with these two models when reading large textual documents on a standard computer monitor. We compare usability of the models and evaluate both user preferences and the effect of the model on performance. Also interesting is the description of the browser and its user interface which we used for the study.

Data Mining

David Stern, Ralf Herbrich and Thore Graepel. Large Scale Online Bayesian Recommendations

Abstract: We present a probabilistic model for generating personalised recommendations of items to users of a web service. The system makes use of content information in the form of user and item meta data in combination with collaborative filtering information from previous user behavior in order to predict the value of an item for a user. Users and items are represented by feature vectors which are mapped into a low-dimensional `trait space' in which similarity is measured in terms of inner products. The model can be trained from different types of feedback in order to learn user-item preferences. Here we present three alternatives: direct observation of an absolute rating each user gives to some items, observation of a binary preference (like/ don't like) and observation of a set of ordinal ratings on a user-specific scale. Efficient inference is achieved by approximate message passing involving a combination of Expectation Propagation (EP) and Variational Message Passing. We also include a dynamics model which allows an items popularity, a user's taste or a user's personal rating scale to drift over time. By using Assumed-Density Filtering (ADF) for training, the model requires only a single pass through the training data. This is an on-line learning algorithm capable of incrementally taking account of new data so the system can immediately reflect the latest user preferences. We evaluate the performance of the algorithm on the MovieLens and Netflix data sets consisting of ~1,000,000 and ~100,000,000 ratings respectively. This demonstrates that training the model using the on-line ADF approach yields state-of-the-art performance with the option of improving performance further if computational resources are available by performing multiple EP passes over the training data.

Sihong Xie, Wei Fan, Jing Peng, Olivier Verscheure and Jiangtao Ren. Latent Space Domain Transfer between High Dimensional Overlapping Distributions

Abstract: Transferring knowledge from one domain to another is challenging due to a number of reasons. Since both conditional and marginal distribution of the training data and test data are non-identical, model trained in one domain, when directly applied to a diffeerent domain, is usually low in accuracy. For many applications with large feature sets, such as text document, sequence data, medical data, image data of different resolutions, etc. two domains usually do not contain exactly the same features, thus introducing large numbers of "missing values" when considered over the union of features from both domains. In other words, its marginal distributions are at most overlapping. In the same time, these problems are usually high dimensional, such as, several thousands of features. Thus, the combination of high dimensionality and missing values make the relationship in conditional probabilities between two domains hard to measure and model. To address these challenges, we propose a framework that first brings the marginal distributions of two domains closer by "filling up" those missing values of disjoint features. Afterwards, it looks for those comparable sub-structures in the "latent-space" as mapped from the expanded feature vector, where both marginal and conditional distribution are similar. With these sub-structures in

latent space, the proposed approach then find common concepts
that are transferable across domains with high probability.
During prediction, unlabeled instances are treated
as "queries", the mostly related labeled instances from
out-domain are retrieved, and the classifcation is made by weighted
voting using retrievd out-domain examples. We formally
show that importing feature values across domains and latent
semantic index can jointly make the distributions of two
related domains easier to measure than in original feature
space, the nearest neighbor method employed to retrieve
related out domain examples is bounded in error when
predicting in-domain examples.

Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang and Ji-Rong Wen. StatSnowball: a Statistical Approach to Extracting Entity Relationships

Abstract: Traditional relation extraction methods require pre-specified relations and relation-specific human-tagged examples. Bootstrapping systems significantly reduce the number of training examples, but they usually apply heuristic-based methods to combine a set of strict hard rules, which will cause limited generalizability and thus low recall. Furthermore, existing bootstrapping methods cannot perform open information extraction (Open IE), which can identify various types of relations without requiring pre-specifications. In this paper, we propose a statistical extraction framework called {\it Statistical Snowball} (StatSnowball), which is a bootstrapping system and can perform both traditional relation extraction and Open IE.

StatSnowball uses the discriminative Markov logic networks (MLNs) and softens hard rules by learning their weights in a maximum likelihood estimate sense. MLN is a general model, and can be configured to perform different levels of relation extraction. In StatSnwoball, pattern selection is performed by solving an ℓ_1-norm penalized maximum likelihood estimation, which enjoys well-founded theories and efficient solvers. We extensively evaluate the performance of StatSnowball in different configurations on both a small but fully labeled data set and large-scale Web data. Empirical results show that StatSnowball can achieve a significantly higher recall without sacrificing the high precision during iterations with a small number of seeds; and the joint inference of MLN can improve the performance. Finally, StatSnowball is efficient and we have developed a working entity relation search engine called {\it Renlifang} based on it.

Guan Hu, Jingyu Zhou and Minyi Guo. A Class-Feature-Centroid Classifier For Text Categorization

Abstract: Automated text categorization is an important technique for many web applications, such as document indexing, document filtering, and cataloging web resources. Many different approaches have been proposed for the automated text categorization problem. Among them, centroid-based approaches have the advantages of short training time and testing time due to its computational efficiency. As a result, centroid-based classifiers have been widely used in many web applications. However, the accuracy of centroid-based classifiers is inferior to SVM, mainly because centroids found during the training process are far from perfect locations.

We design a fast Class-Feature-Centroid (CFC) classifier for multi-class, single-label text categorization. In CFC, a centroid is built from two important class features: inter-class term distribution and inner-class term distribution. CFC proposes a novel combination of these features and employs a denormalized

cosine measure to calculating the similarity between a text vector and a centroid. Experiments on the Reuters-21578 corpus and 20-newsgroup email collection show that CFC consistently outperforms the state-of-the-art SVM classifiers on both micro-F1 and macro-F1 scores. Particularly, CFC is more effective and robust than SVM when the training data is sparse.

Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka. Measuring the Similarity between Implicit Semantic Relations from the Web

Abstract: Measuring the similarity between semantic relations that hold among entities
is an important and necessary step in various Web related tasks such as relation extraction, information retrieval and analogy detection.
For example, consider the case in which a person knows a pair of entities (e.g. \textit{Google, YouTube}),
between which a particular relation holds (e.g. acquisition).
The person is interested in retrieving other such pairs
with similar relations (e.g. \textit{Microsoft, Powerset}).
Existing keyword-based search engines cannot be applied directly in this case because, in keyword-based search, the goal is to retrieve documents that are relevant to the words used in a query --
not necessarily to the relations implied by a pair of words.
We propose a relational similarity measure, using a Web search engine, to compute the
similarity between semantic relations implied by two pairs of words.
Our method has three components:
representing the various semantic relations that exist between a pair of words using automatically extracted lexical patterns,
clustering the extracted lexical patterns to identify the different patterns that express a particular
semantic relation,
and measuring the similarity between semantic relations using a metric learning approach.
We evaluate the proposed method in two tasks: classifying semantic relations between named entities, and solving word-analogy questions.
The proposed method outperforms all baselines in a relation classification task with a statistically significant average precision score of 0.74.
Moreover, it reduces the time take by Latent Relational Analysis to process 374 word-analogy
questions from 9 days to less than 6 hours, with a SAT score of 51%.

Jiang-Ming Yang, Rui Cai, Yida Wang, Jun Zhu, Lei Zhang and Wei-Ying Ma. Incorporating Site-Level Knowledge to Extract Structured Data from Web Forums

Abstract: Web forum has become an important data resource for many Web applications, but extracting structured data from unstructured Web forum pages is still a challenging task due to both complex page layout designs and unrestricted user created posts. In this paper, we study the problem of structured data extraction from various Web forum sites. Our target is to find a solution as general as possible to extract structured data such as post title, post author, post time, and post content from any forum site. In contrast to most existing information extraction methods which only leverage the knowledge inside an individual page, we incorporate both the page-level and site-level knowledge and employ Markov logic networks (MLNs) to effectively integrate all useful evidences by learning their importance automatically. The site-level knowledge includes (1) the linkages among different object pages such as list pages and post pages, and (2) the interrelationships of pages belonging to one same object. The experimental results on 20 forums show very encouraging performance of information extraction, and

demonstrate the generalization ability of the proposed approach on various forums. We also show that the performance is limited if only page-level knowledge is used, while incorporating the site-level knowledge both precision and recall can be significantly improved.

Ossama Abdelhamid, Behshad Behzadi, Stefan Christoph and Monika Henzinger. Detecting The Origin Of Text Segments Efficiently

Abstract: In the origin detection problem an algorithm is given a set S of documents, ordered by creation time, and a query document D. It needs to output for every consecutive sequence of k alphanumeric terms in D the earliest document in S in which the sequence appeared (if such a document exists.). Algorithms for the origin detection problem can, for example, be used to detect the ``origin'' of text segments in D and thus to detect novel content in D. They can also find the document from which the author of D has copied the most (or show that D is mostly original.).

%We concentrate on solutions that use only a fixed amount of memory. We propose novel algorithms for this problem and evaluate them together with a large
number of previously published algorithms. Our results show that (1) detecting the origin of text segments efficiently can be done with very high accuracy even when the space used is less than 1$\%$ of the size of the documents in S, (2) the precision degrades smoothly with the amount of available space, (3) various estimation techniques can be used to increase the performance of the algorithms.

Yue Lu, ChengXiang Zhai and Neel Sundaresan. Rated Aspect Summarization of Short Comments

Abstract: Web 2.0 technologies have enabled more and more people to freely comment on different kinds of entities (e.g. sellers, products, services). The large scale of information poses the need and challenge of automatic summarization. In many cases, each of the user generated short comments comes with an overall rating. In this paper, we study the problem of generating a ``rated aspect summarization'' of short comments, which is a decomposed view of the overall ratings for the major aspects so that a user could gain different perspectives towards the target entity. We formally define the problem and decompose the solution into three steps. We demonstrate the effectiveness of our methods by using eBay sellers' feedback comments. We also quantitatively evaluate each step of our methods and study how human agree on such summarization task. The proposed methods are quite general and can be used to generate rated aspect summary given any collection of short comments each associated with an overall rating.

Ziv Bar-Yossef and Maxim Gurevich. Estimating the ImpressionRank of Web Pages

Abstract: The ImpressionRank of a web page (or, more generally, of a web site) is the number of times users viewed the page while browsing search results. ImpressionRank captures the visibility of pages and sites in search engines and is thus an important measure, which is of interest to web site owners, competitors, market analysts, and end users.

All previous approaches to estimating the ImpressionRank of a page rely on privileged access to private data sources, like the search engine's query log. In this paper we present the first {\em external} algorithm for estimating the ImpressionRank of a web page. This algorithm relies on access to three public data sources: the search engine, the query suggestion service of the search engine, and the

web. In addition, the algorithm is {\em local} and uses modest resources. It can therefore be used by almost any party to estimate the ImpressionRank of any page on any search engine. Empirical analysis of the algorithm on the Google and Yahoo!\ search engines indicates that it is accurate and provides interesting observations on sites and search queries.

Surajit Chaudhuri, Venkatesh Ganti and Dong Xin. Mining the Web to Facilitate Fast and Accurate Approximate Match

Abstract: Tasks relying on recognizing entities have recently received significant attention in the literature. Many such tasks assume the existence of reference entity tables. In this paper, we consider the problem of determining whether a candidate string approximately matches with a reference entity. This problem is important for
extracting named entities such as products or locations from a reference entity table, or matching entity entries across heterogenous sources. Prior approaches have relied on string-based similarity which only compare a candidate string and an entity it matches with. In this paper, we observe that considering such evidence across multiple documents significantly improves the accuracy of matching. We develop efficient techniques which exploit web search engines to facilitate approximate matching in the context of our proposed similarity functions. In an extensive experimental evaluation, we demonstrate the accuracy and efficiency of our techniques.

jong wook kim, K. Selcuk Candan and Junichi Tatemura. Efficient Overlap and Content Reuse Detection in Blogs and Online News Articles

Abstract: The use of blogs to track and comment on real world (political, news, entertainment) events is growing. Similarly, as more individuals start relying on the Web as their primary information source and as more traditional media outlets try reaching consumers through alternative venues, the number of news sites on the Web is also continuously increasing. Content-reuse, whether in the form of extensive quotations
or content borrowing across media outlets, is very common in blogs and news entries outlets tracking the same real-world event. Knowledge about which web entries re-use content from which others can be an effective asset when organizing these entries for presentation. On the other hand, this knowledge is not cheap to acquire: considering the size of the related space web entries, it is essential that the techniques developed for identifying re-use are fast and scalable. Furthermore, the dynamic nature of blog and news entries necessitates incremental processing for reuse detection. In this paper, we develop a novel qSign algorithm that effciently and effectively analyze the blogosphere for quotation and reuse identification. Experiment results show that with qSign processing time gains from 10X to 100X are possible while maintaining reuse detection rates of upto 90%. Furthermore, processing time gains can be pushed multiple orders of magnitude (from 100X to 1000X) for 70% recall.

deepak agarwal, Bee-Chung Chen and Pradheep Elango. Spatio-Temporal Models for Estimating Click-through Rate

Abstract: We propose novel spatio-temporal models to estimate click-through rates in the context of content recommendation. We track article CTR at a fixed location over time through a dynamic Gamma-Poisson model and combine information from correlated locations through dynamic linear regressions, significantly improving on per-location model. Our models adjust for user fatigue through an exponential tilt to the first-view CTR (probability of click on first article exposure) that is based only on user-specific repeat-exposure features. We illustrate our approach on data obtained from a module (Today Module) published regularly on Yahoo! Front Page and demonstrate significant improvement over

commonly used baseline methods. Large scale simulation experiments to study the performance of our models under different scenarios provide encouraging results. Throughout, all modeling assumptions are validated via rigorous exploratory data analysis.

Lei Tang, Suju Rajan and Vijay Narayanan. Large Scale Multi-Label Classification via MetaLabeler

Abstract: The explosion of online content has made the management of such content non-trivial. Web-related tasks such as web page categorization, news filtering, query categorization, tag recommendation, etc. often involve the construction of multi-label categorization systems on a large scale. Existing multi-label classification methods either do not scale or have unsatisfactory performance. In this work, we propose MetaLabeler to automatically determine the relevant set of labels for each instance without intensive human involvement or expensive cross-validation. Extensive experiments conducted on benchmark data show that the MetaLabeler tends to outperform existing methods. Moreover, MetaLabeler scales to millions of multi-labeled instances and can be deployed easily. This enables us to apply the MetaLabeler to a large scale query categorization problem in Yahoo!, yielding a significant improvement in performance.

Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha and Yong Yu. Summarization through Structure Learning with Diversity, Coverage and Balance

Abstract: Document summarization has played an ever more important role with the exponential growth of documents on the web. Many supervised and unsupervised approaches have been proposed to extract summaries from documents. However, these approaches seldom consider summary diversity, coverage, and balance issues which to a large extent determine the quality of summaries. In this paper we consider extract-based summarization with emphasis placed on three requirements: 1) diversity in summarization which seeks to reduce redundancy among sentences in the summary; 2) sufficient coverage which focuses on avoiding loss of key information of the document in the summary; and 3) balance which demands a equal amount of information about different aspects of a document in the summary. We formulate the extract-based summarization problem as learning a mapping from a set of sentences of a given document to a subset of the sentences that satisfies the above three requirements. The mapping is learned by incorporating several constraints in a structured learning framework to enhance diversity, coverage and balance of the summaries. Furthermore, we explore a graph structure of the output variable in the structure learning problem and employ structured SVM for its solution. Experiments on the DUC2001 data sets demonstrate significant improvement of performance in terms of the F1 and ROUGE metrics.

Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra and Alexandros Ntoulas. Releasing Search Queries and Clicks Privately

Abstract: The question of how to publish an anonymized search log was brought to the forefront by a well-intentioned, but privacy-unaware AOL search log release. Since then a series of ad-hoc techniques have been proposed in the literature, though none are known to be provably private. In this paper, we take a major step towards a solution: we show how queries, clicks and their associated perturbed counts can be published in a manner that rigorously preserves privacy. Our algorithm is decidedly simple to state, but non-trivial to analyze. On the opposite side of privacy is the question of whether the data we can safely publish is of any use. Our findings offer a glimmer of hope: we demonstrate that a non-negligible fraction of queries and clicks can indeed be safely published via a

collection of experiments on a real search log. In addition, we select an application, finding similar queries, and show that the similar queries discovered on the original data resemble those found on the perturbed data.

Purnamrita Sarkar and Andrew Moore. Fast Dynamic Reranking in Large Graphs

Abstract: In this paper we consider the problem of re-ranking search results
by incorporating user feedback. We present a graph theoretic measure for
discriminating
irrelevant results from relevant results using a few labeled examples provided by
the user.
The key intuition is that nodes relatively closer (in graph topology) to the relevant
nodes
than the irrelevant nodes are more likely to be relevant.
We present a simple sampling algorithm to evaluate this measure at specific nodes
of interest, and an efficient
branch and bound algorithm to compute the top k nodes from the entire graph
under this measure.
On quantifiable prediction tasks the introduced measure outperforms other
diffusion-based proximity measures which take only the positive relevance feedback
into account. On the entity-relation graph built from the authors and papers of
the entire DBLP citation corpus (1.4 million nodes and 2.2 million edges) our
branch and bound
algorithm takes about 1.5 seconds to retrieve the top 10 nodes w.r.t. this
measure with 10 labeled nodes.

Fan Guo, Chao Liu, Tom Minka, Yi-Min Wang and Christos Faloutsos. Click Chain Model in Web Search

Abstract: Given a terabyte click log, can we build an efficient and effective click model? It is commonly believed that web search click logs are a gold mine for search business, because they reflect users' preference over web documents presented by the search engine. Click models provide a principled approach to inferring user-perceived relevance of web documents, which can be leveraged in numerous applications in search businesses. Due to the huge volume of click data, scalability is a must.

We present the click chain model (CCM), which is based on a solid, Bayesian framework. It is both scalable and incremental, perfectly meeting the computational challenges imposed by the voluminous click logs that constantly grow. We conduct a reproducible experimental study on a data set containing 8.8 million query sessions obtained in July 2008 from a commercial search engine. CCM consistently outperforms two state-of-the-art competitors in a number of metrics, with over 12% better log-likelihood, more than 6% improvement in perplexity and 10% improvement in the prediction quality of the first and the last clicked position.

Internet Monetization

Ashish Goel and Kamesh Munagala. Hybrid Keyword Search Auctions

Abstract: Search auctions have become a dominant source of revenue generation on the Internet. Such auctions have typically used per-click bidding and pricing. We propose the use of hybrid auctions where an advertiser can make a per-impression as well as a per-click bid, and the auctioneer then chooses one of the two as the pricing mechanism. We assume that the advertiser and the auctioneer both have separate beliefs (called priors) on the click-probability of an advertisement. We first prove that the hybrid auction is truthful, assuming that the advertisers are risk-neutral. We then show that this auction is superior to the existing per-click auction in multiple ways:

1) We show that risk-seeking advertisers will choose only a per-impression bid whereas risk-averse advertisers will choose only a per-click bid, and argue that both kind of advertisers arise naturally. Hence, the ability to bid in a hybrid fashion is important to account for the risk characteristics of the advertisers.

2) For obscure keywords, the auctioneer is unlikely to have a very sharp prior on the click-probabilities. In such situations, we show that having the extra information from the advertisers in the form of a per-impression bid can result in significantly higher revenue.

3) An advertiser who believes that its click-probability is much higher than the auctioneer's estimate can use per-impression bids to correct the auctioneer's prior without incurring any extra cost.

4) The hybrid auction can allow the advertiser and auctioneer to implement complex dynamic programming strategies to deal with the uncertainty in the click-probability using the same basic auction. The per-click and per-impression bidding schemes can only be used to implement two extreme cases of these strategies.

As Internet commerce matures, we need more sophisticated pricing models to exploit all the information held by each of the participants. We believe that hybrid auctions could be an important step in this direction. The hybrid auction easily extends to multiple slots, and is also applicable to scenarios where the hybrid bidding is per-impression and per-action (i.e. CPM and CPA), or per-click and per-action (i.e. CPC and CPA).

Gagan Aggarwal, S Muthukrishnan, David Pal and Martin Pal. General Auction Mechanism for Search Advertising

Abstract: In sponsored search, a number of advertising slots is available on a search results page, and have to be allocated among a set of advertisers competing to display an ad on the page. This gives rise to a bipartite matching market that is typically cleared by the way of an automated auction. Several auction mechanisms have been proposed, with variants of the Generalized Second Price (GSP) being widely used in practice.

There is a rich body of work on bipartite matching markets that builds upon the stable marriage model of Gale and Shapley and the assignment model of Shapley and Shubik. This line of research offers deep insights into the structure of stable outcomes in such markets and their incentive properties.

In this paper, we model advertising auctions in terms of an assignment model with linear utilities, extended with bidder and item specific maximum and minimum prices. Auction mechanisms like the commonly used GSP or the well-known Vickrey-Clarke-Groves (VCG) can be interpreted as simply computing a \emph{bidder-optimal stable matching} in this model, for a suitably defined set of bidder preferences, but our model includes much richer bidders and preferences. We prove that in our model the existence of a stable matching is guaranteed, and under a non-degeneracy assumption a bidder-optimal stable matching exists as well. We give a fast algorithm to find such matching in polynomial time, and use it to design truthful mechanism that generalizes GSP, is truthful for profit-maximizing bidders, correctly implements features like bidder-specific minimum prices and position-specific bids, and works for rich mixtures of bidders and preferences.
Our main technical contributions are the existence of bidder-optimal matchings and (group) strategyproofness of the resulting mechanism, and are proved by induction on the progress of the matching algorithm.

Jun Yan, Ning Liu, Gang Wang, wen zhang, yun jiang and Zheng Chen. How much the Behavioral Targeting can Help Online Advertising?

Abstract: Behavioral Targeting (BT) attempts to deliver the most relevant advertisements to the most interested audiences, and is playing an increasingly important role in online advertising market. However, there have been not any public works investigating on how much the BT can truly help online advertising in commercial search engines? To answer this question, in this paper we provide an empirical study on the ads click-through log collected from a commercial search engine. From the comprehensively experimental results on the sponsored search log of a commercial search engine over a period of seven days, we can draw three important conclusions: (1) Users who clicked the same ad will truly have similar behaviors on the Web; (2) The Click-Through Rate (CTR) of an ad can be averagely improved as high as 670% by properly segmenting users for behavioral targeted advertising; (3) Using the short term user behaviors to represent users is more effective than using the long term user behaviors for BT. The statistical t-test verifies that all conclusions drawn in the paper are statistically significant. To our best knowledge, this work is the first empirical study for BT on real world ads click-through log in academia.

Arpita Ghosh, Benjamin Rubinstein, Sergei Vassilvitskii and Martin Zinkevich. Adaptive Bidding for Display Advertising

Abstract: Motivated by the emergence of auction-based marketplaces for display ads such as the Right Media Exchange, we study the design of a bidding agent that implements a display advertising campaign by bidding in such a marketplace. The bidding agent must acquire a given number of impressions with a given target spend, when the highest external bid in the marketplace is drawn from an {\em unknown} distribution $\cP$. The quantity and spend constraints arise from the fact that display ads are usually sold on a CPM basis. We consider both the full information setting, where the winning price in each auction is announced publicly, and the partially observable setting where only the winner obtains information about the distribution; these differ in the penalty incurred by the agent while attempting to learn the distribution. We provide algorithms for both settings, and prove performance guarantees using bounds on uniform closeness from statistics, and techniques from online learning. We experimentally evaluate these algorithms: both algorithms perform very well with respect to both target quantity and spend; further, our algorithm for the partially observable case performs nearly as well as

that for the fully observable setting despite the higher penalty incurred during learning.

Vahab Mirrokni, Eyal Even-Dar, Yishay Mansour, S Muthukrishnan and Uri Nadav. Bid Optimization for Broad Match Ad Auctions

Abstract: Ad auctions support the ``broad match'' that
allows an advertiser to target a large number of queries by bidding only
on a limited number of queries as broad match.
While giving more expressiveness to advertisers, this feature makes it challenging
for the advertisers to find bidding strategies to optimize their returns:
choosing to bid on a query as a broad match because it provides
high profit results in one bidding for related queries which may yield low or even
negative profits.

We abstract and study the complexity of the {\em bid optimization problem} which is to
determine an advertiser's bids on
a subset of keywords (possibly using broad match) so that her profit is maximized.
In the query language model when the advertiser is allowed to bid on all queries as broad match, we present
an linear programming (LP)-based polynomial-time algorithm for the bid optimization problem.
In the model in which an advertiser can only bid on keywords, ie.,
a subset of keywords as an exact or broad match, we show that this problem is not approximable
within any reasonable approximation factor unless P=NP. To deal with this hardness result,
we present a constant-factor approximation when the optimal profit significantly exceeds the cost. This algorithm is based on rounding
a natural LP formulation of the problem.
Finally, we study a budgeted variant of the problem, and show that in the query language model, one
can find two budget constrained ad campaigns in polynomial time that implement the optimal bidding strategy.
Our results are the first to address bid optimization under the broad match feature which is
common in ad auctions.

Thomas Meinl and Benjamin Blau. Web Service Derivatives

Abstract: Web service development and usage has shifted from simple information processing services to high-value business services that are crucial to productivity and success. In order to deal with an increasing risk of unavailability or failure of mission-critical Web services we argue the need for advanced reservation of services in the form of derivatives.

The contribution of this paper is twofold: First we provide an abstract model of a market design that enables the trade of derivatives for mission-critical Web services. Our model satisfies requirements that result from service characteristics such as intangibility and the impossibility to inventor services in order to meet fluctuating demand. It comprehends principles from models of incomplete markets such as the absence of a tradeable underlying and consistent arbitrage-free derivative pricing.

Furthermore we provide an architecture for a Web service market that implements our model and describes the strategy space and interaction of market participants in the trading process of service derivatives. We compare the underlying pricing processes to existing derivative models in energy exchanges, discuss eventual shortcomings, and apply Wavelets to analyze actual data and extract long- and short-term trends.

Performance, Scalability and Availability

Zakaria Al-Qudah, Hussein Alzoubi, Mark Allman, Michael Rabinovich and Vincenzo Liberatore. Efficient Application Placement in a Dynamic Hosting Platform

Abstract: Web hosting providers are increasingly looking into dynamic hosting to reduce costs and improve the performance of their platforms. Instead of provisioning fixed resources to each customer, dynamic hosting maintains a variable number of application instances to satisfy current demand. While existing research in this area has mostly focused on the algorithms that decide on the number and location of application instances, we address the problem of efficient enactment of these decisions once they are made. We propose a new approach to application placement and experimentally show that it dramatically reduces the cost of application placement, which in turn improves the end-to-end agility of the hosting platform in reacting to demand changes.

Toyotaro Suzumura, Michiaki Tatsubori, Scott Trent, Akihiko Tozawa and Tamiya Onodera. Highly Scalable Web Applications with Zero-Copy Data Transfer

Abstract: The performance of server side applications is becoming increasingly important as more applications exploit the web application model. Extensive work has been done to improve the performance of individual software components such as web servers and programming language runtimes. This paper describes a novel approach to boost web application performance by improving interprocess communication between the programming language runtime and operating system. The approach reduces redundant processing for memory copying and the context switch overhead between users space and kernel space by exploiting the zero-copy data transfer methodology such as the sendfile system call. In order to transparently utilize this optimization feature with existing web applications, we propose an enhancement of the PHP runtime, FastCGI protocol, and web server. Our proposed approach achieves a 126% performance improvement with micro-benchmarks, and a 22% performance improvement for the standard web benchmark, SPECweb2005.

Jeffrey Erman, Alexandre Gerber, Oliver Spatscheck, Dan Pei and MohammadTaghi Hajiaghayi. Network Aware Forward Caching

Abstract: This paper proposes and evaluates a Network Aware Forward Caching approach for determining the optimal deployment strategy of forward caches to a network. A key advantage of this approach is that we can reduce the costs associated with forward caching to maximize the benefit obtained from their deployment. We show in our simulation that a 37% increase to net benefits could be achieved over the standard method of full cache deployment to cache all POPs traffic. In addition, we show that this maximal point occurs when only 68% of the total traffic is cached.
Another contribution of this paper is the analysis we use to motive and evaluate this problem. We characterize the Internet traffic of 100K subscribers of a US residential broadband provider. We use both layer 4 and layer 7 analysis to investigate the traffic volumes of the flows as well as study the general characteristics of the applications used. We show that HTTP is a dominant protocol and account for 68% of the total downstream traffic and that 34% of that traffic is multimedia. In addition, we show that multimedia content using HTTP exhibits a 83% annualized growth rate and other HTTP

traffic has a 53% growth rate versus the 26% over all annual growth rate of broadband traffic. This shows that HTTP traffic will become ever more dominent and increase the potential caching opportunities. Furthermore, we characterize the core backbone traffic of this broadband provider to measure the efficiency content and traffic is delivered. We find that CDN traffic is much more efficient than P2P content and that there is large skew in the Air Miles between POP in a typical network. Our findings show that there are many opportunties in broadband provider networks to optimize how traffic is delivered and cached.

Zakaria Al-Qudah, Seungjoon Lee, Michael Rabinovich, Oliver Spatscheck and Kobus van der Merwe. Anycast-Aware Transport for Content Delivery Networks

Abstract: Anycast-based content delivery networks (CDNs) have many properties that make them ideal for the large scale distribution of content on the Internet. However, because routing changes can result in a change of the endpoint that terminates the TCP session, TCP session disruption remains a concern for anycast CDNs, especially for large file down-loads. In this paper we demonstrate that this problem does not require any complex solutions. In particular, we present the design of a simple, yet efficient, mechanism to handle session disruptions due to endpoint changes. With our mechanism, a client can continue the download of the content from the point at which it was before the endpoint change. Furthermore, CDN servers purge the TCP connection state quickly to handle frequent switching with low system overhead. We demonstrate experimentally the effectiveness of our proposed mechanism and show that more complex mechanisms are not required. Specifically, we find that our mechanism maintains high download throughput even with a reasonably high rate of endpoint switching, which is attractive for load balancing scenarios. Moreover, our results show that edge servers can purge TCP connection state after a single timeout-triggered retransmission without any tangible impact on ongoing connections. Besides improving server performance, this behavior improves the resiliency of the CDN to certain denial of service attacks.

Rich Media

Alberto Messina and Maurizio Montagnuolo. A Generalised Cross-Modal Clustering Method Applied to Multimedia News Semantic Indexing and Retrieval

Abstract: The evolution of Web services has enabled the distribution of informative content through dynamic media such as RSS feeds. In addition, the availability of the same informative content in the form of digital multimedia data has dramatically increased. Robust solutions for semantic aggregation of heterogeneous data streams are needed to efficiently access desired information from this variety of information sources. To this end, we present a novel approach for cross-media information sources aggregation, and describe a prototype system implementing this approach. The prototype adopts online newspaper articles and TV newscasts as information sources, to deliver a service made up of items including both contributions. Extensive experiments prove the effectiveness of the proposed approach in a real-world business context.

Reinier H. van Leuken, Lluís Garcia, Ximena Olivares and Roelof van Zwol. Visual diversification of image search results

Abstract: Due to the reliance on the textual information associated with an image, image search engines on the Web lack the discriminative power to deliver visually diverse search results. The textual descriptions are key to retrieve relevant results for a given user query, but at the same time provide little information about the rich image content.

In this paper we investigate three methods for visual diversification of image search results. The methods deploy lightweight clustering techniques in combination with a dynamic weighting function of the visual features, to best capture the discriminative aspects of the resulting set of images that is retrieved. A representative image is selected from each cluster, which together form a diverse result set.

Based on a performance evaluation we find that the outcome of the methods closely resembles human perception of diversity, which was established in an extensive clustering experiment carried out by human assessors.

Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang and Hong-Jiang Zhang. Tag Ranking

Abstract: Social media sharing web sites like Flickr allow users to annotate images with free tags, which significantly facilitate Web image search and organization. However, the tags associated with an image generally are in a random order without any importance or relevance information, which limits the effectiveness of these tags in search and other applications. In this paper, we propose a tag ranking scheme, aiming to automatically rank the tags associated with a given image according to their relevance to the image content. We first estimate initial relevance scores for the tags based on probability density estimation, and then perform a random walk over a tag similarity graph to refine the relevance scores. Experimental results on a 50, 000 Flickr photo collection show that the proposed tag ranking method is both effective and efficient. We also apply tag ranking into three applications: (1) tag-based image search, (2) tag recommendation, and (3) group recommendation, which demonstrates that the proposed tag ranking approach really boosts the performances of social-tagging related applications.

Lyndon Kennedy and Mor Naaman. Less Talk, More Rock: Automated Organization of Community-Contributed Collections of Concert Videos

Abstract: We describe a system for synchronization and organization of user-contributed content from live music events. We start with a set of short video clips taken at a single event by multiple contributors, who were using a varied set of capture devices. Using audio ngerprints, we synchronize these clips such that overlapping clips can be displayed simultaneously. Furthermore, we use the timing and link structure generated by the synchronization algorithm to improve the ndability and representation of the event content, including identifying key moments of interest and descriptive text for important captured segments of the show. We also identify the preferred audio track when multiple clips overlap. We thus create a much improved representation of the event that builds on the automatic content match. Our work demonstrates important principles in the use of content analysis techniques for social media content on the Web, and applies those principles in the domain of live music capture.

Lei Wu, Linjun Yang, Nenghai YU and Xian-Sheng Hua. Learning to Tag

Abstract: Social tagging provides valuable and crucial information for large-scale web image retrieval. It is ontology-free and easy to obtain; however, noisy tags frequently appear, and users typically will not tag all semantic objects in the image, which is also called semantic loss. To avoid noises and compensate for the semantic loss, tag recommendation is proposed in literature. However, current recommendation simply ranks the related tags based on the single modality of tag co-occurrence on the whole dataset, which ignores other modalities, such as visual correlation. This paper proposes a multi-modality recommendation based on both tag and visual correlation, and formulates the tag recommendation as a learning problem. Each modality is used to generate a ranking feature, and Rankboost algorithm is applied to learn an optimal combination of these ranking features from different modalities. Experiments on Flickr data demonstrate the effectiveness of this learning-based multi-modality recommendation strategy.

Search

Xing Yi, Hema Raghavan and Chris Leggetter. Discover Users' Specific Geo Intention in Web Search

Abstract: Discovering users' specific and implicit geographic intention in web search can greatly help satisfy users' information needs. We build a geo intent analysis system that learns a model from large scale web-search logs for this discovery with minimal supervision. We build a city language model, which is a probabilistic representation of the language surrounding the mention of a city in web queries. We use several features derived from these language models to: (1) identify users' implicit geo intent and pinpoint the city corresponding to this intent (2)determine whether the geo-intent is localized around the users' current geographic location. (3) predict cities for queries that have a mention of an entity that is located in a specific place. Experimental results demonstrate the effectiveness of using features derived from the city language model. We find that (1) the system has over 90% precision and more than 74% accuracy for the task of detecting users' implicit city level geo intent. (2) the system achieves more than 96% accuracy in determining whether implicit geo queries are local geo queries, neighbor region geo queries or none-of these (3) the city language model can effectively retrieve cities in location-specific queries with high precision(88%) and recall(74%); human evaluation results show that the language model predicts city labels for location-specific queries with high accuracy (84.5%).

Xiangfu Meng, Z. M. Ma and Li Yan. Answering Approximate Queries over Autonomous Web Databases

Abstract: To deal with the problem of empty or too little answers returned from a Web database in response to a user query, this paper proposes a novel approach to provide relevant and ranked query results. Based on the user original query, we speculate how much the user cares about each specified attribute and assign a corresponding weight to it. This original query is then rewritten as an approximate query by relaxing the query criteria range. The relaxation order of all specified attributes and the relaxed degree on each specified attribute are varied with the attribute weights. For the approximate query results, we generate users' contextual preferences from database workload and use them to create a priori orders of tuples in an off-line preprocessing step. Only a few representative orders are saved, each corresponding to a set of contexts. Then, these orders and associated contexts are used at query time to expeditiously provide ranked answers. Results of a preliminary user study demonstrate that our query relaxation and results ranking methods can capture the user's preferences effectively. The efficiency and effectiveness of our approach is also demonstrated by experimental result.

Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen and Hang Li. Towards Context-Aware Search by Learning A Very Large Variable Length Hidden Markov Model from Search Logs

Abstract: Capturing the context of a user's query from the previous queries and clicks in the same session leads to better understanding of the user's information need. A context-aware approach to document re-ranking, query suggestion, and URL recommendation may improve users' search experience substantially. In this paper, we propose a general approach to context-aware search. To capture contexts of queries, we learn a variable length Hidden Markov Model (vlHMM) from search sessions extracted from log data. Although the mathematical model is intuitive, how to learn a large vlHMM with millions of states from hundreds of millions of search sessions poses a grand challenge. We develop a strategy for parameter initialization in vlHMM learning which can greatly reduce the number of parameters to be estimated in practice. We also devise a method for distributed

vlHMM learning under the map-reduce model. We test our approach on a real data set consisting of 1.8 billion queries, 2.6 billion clicks, and 840 million search sessions, and evaluate the effectiveness of the vlHMM learned from the real data on three search applications: document re-ranking, query suggestion, and URL recommendation. The experimental results clearly show that our context-aware approach is both effective and efficient.

Eustache Diemert and Gilles Vandelle. Unsupervised Query Categorization using Automatically-Built Concept Graphs

Abstract: Automatic categorization of user queries is an important component of general purpose (Web) search engines, particularly for triggering rich, query-specific content and sponsored links. We propose an unsupervised learning scheme that reduces dramatically the cost of setting up and maintaining such a categorizer, while retaining good categorization power. The model is stored as a graph of concepts where graph edges represent the cross-reference between the concepts. Concepts and relations are extracted from query logs by an offline Web mining process, which uses a search engine as a powerful summarizer for building a concept graph. Empirical evaluation indicates that the system compares favorably on publicly available data sets (such as KDD Cup 2005) as well as on portions of the current query stream of Yahoo! Search, where it is already changing the experience of millions of Web search users.

Jian Hu, gang wang, Fred Lochovsky and Zheng Chen. Understanding User's Query Intent with Wikipedia

Abstract: Understanding the intent behind a user's query can help search engine to automatically route the query to some corresponding vertical search engines to obtain particularly relevant contents, thus, greatly improving user satisfaction. There are three major challenges to the query intent classification problem: (1) Intent representation; (2) Domain coverage and (3) Semantic interpretation. Current approaches to predict the user's intent mainly utilize machine learning techniques. However, it is difficult and often requires much human efforts to meet all these challenges by the statistical machine learning approaches. In this paper, we propose a general methodology to the problem of query intent classification. With very little human effort, our method can discover large quantities of intent concepts by leveraging Wikipedia, one of the best human knowledge base. The Wikipedia concepts are used as the intent representation space, thus, each intent domain is represented as a set of Wikipedia articles and categories. The intent of any input query is identified through mapping the query into the Wikipedia representation space. Compared with previous approaches, our proposed method can achieve much better coverage to classify queries in an intent domain even through the number of seed intent examples is very small. Moreover, the method is very general and can be easily applied to various intent domains. We demonstrate the effectiveness of this method in three different applications, i.e., travel, job, and person name. In each of the three cases, only a couple of seed intent queries are provided. We perform the quantitative evaluations in comparison with two baseline methods, and the experimental results shows that our method significantly outperforms other methods in each intent domain.

Sreenivas Gollapudi and Aneesh Sharma. An Axiomatic Approach to Result Diversification

Abstract: Understanding user intent is key to designing an effective ranking system in a search engine. In the absence of any explicit knowledge of user intent, search engines want to diversify results to improve user satisfaction. In such a setting, the probability ranking principle-based approach of presenting the most relevant results on top can be sub-optimal, and hence the search engine would like to trade-off relevance for diversity in the results.

In an analogy to prior work on ranking and clustering systems, we use the axiomatic approach to characterize and design diversification systems. We develop a set of natural axioms that a diversification system is expected to satisfy, and show that no diversification function can satisfy all the axioms simultaneously. We illustrate the use of the axiomatic framework by providing three example diversification objectives that satisfy different subsets of the axioms. We also uncover a rich link to the facility dispersion problem that results in algorithms for a number of diversification objectives. Finally, we propose an evalutation methodology to characterize the objectives and the underlying axioms. We conduct a large scale evaluation of our objectives based on two data sets: a data set derived from the Wikipedia disambiguation pages and a product database.

Flavio Chierichetti, Ravi Kumar and Prabhakar Raghavan. Compressed web indexes

Abstract: Web search engines use indexes to efficiently retrieve pages containing specified query terms, as well as pages linking to specified pages. The problem of compressed indexes that permit such fast retrieval has a long history. We consider the problem: assuming that the terms in (or links to) a page are generated from a probability distribution, how well compactly can we build such indexes that allow fast retrieval? Of particular interest is the case when the probability distribution is Zipfian (or similar power laws), since these are the distributions that arise on the web.

We obtain sharp bounds on the space requirement of Boolean indexes for text documents that follow Zipf's law. In the process we develop a general technique that applies to any probability distribution, not necessarily a power law. Our bounds lead to quantitative versions of rules of thumb that are folklore in indexing. Our experiments on several document collections show that in reality, the distributions of terms appear to follow a double-Pareto law rather than Zipf's law. Index sizes observed in the experiments conform well to our theoretical predictions.

Andrei Broder, Flavio Chierichetti, Vanja Josifovski, Ravi Kumar, Sandeep Pandey and Sergei Vassilvitskii. Nearest-Neighbor Caching for Content-Match Applications

Abstract: Motivated by contextual advertising systems and other web applications involving efficiency-accuracy tradeoffs, we study similarity caching. Here, a cache hit is said to occur if the requested item is similar but not necessarily equal to some cached item. We study two objectives that dictate the efficiency-accuracy tradeoff and provide our caching policies for these objectives. By conducting extensive experiments on real data we show similarity caching can significantly improve the efficiency of contextual advertising systems, with minimal impact on accuracy. Inspired by the above, we propose a simple generative model that embodies two fundamental characteristics of page requests arriving to advertising systems, namely, long-range dependences and similarities. We provide theoretical bounds on the gains of similarity caching in this model and demonstrate these gains empirically by fitting the actual data to the model.

Paul Bennett, Max Chickering and Anton Mityagin. Learning Consensus Opinion:Mining Data from a Labeling Game

Abstract: In this paper, we consider the challenge of how to identify the consensus opinion of a set of users as to how the results for a query should be ranked. Once consensus rankings are identified for a set of queries, these rankings can serve for both evaluation and training of retrieval and learning systems. We present a novel approach to

collecting user preferences over image-search results: we use a collaborative game in which players are rewarded for agreeing on which image result is best for a query. Our approach is distinct from other labeling games because we are able to elicit directly the preferences of interest with respect to image queries extracted from query logs. As a source of relevance judgments, this data provides a useful complement to click data. Furthermore, it is free of positional biases and does not carry the risk of frustrating users with non-relevant results associated with proposed mechanisms for debiasing clicks. We describe data collected over 35 days from a deployed version of this game that amounts to about 19 million expressed preferences between pairs. Finally, we present several approaches to modeling this data in order to extract the consensus rankings from the preferences and better sort the search results for targeted queries.

Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler, Lance Riedel and Jeffrey Yuan. Online Expansion of Rare Queries for Sponsored Search

Abstract: Sponsored search systems are tasked with matching queries to relevant advertisements. The current state-of-the-art matching algorithms expand the user's query using a variety of external resources, such as Web search results. While these expansion-based algorithms are highly effective, they are largely inefficient and cannot be applied in real-time. In practice, such algorithms are applied offline to popular queries, with the results of the expensive operations cached for fast access at query time. In this paper, we describe an efficient and effective approach for matching ads against rare queries that were not processed offline. The approach builds an expanded query representation by leveraging offline processing done for related popular queries. Our experimental results show that our approach significantly improves the effectiveness of advertising on rare queries with only a negligible increase in computational cost.

Xuerui Wang, Andrei Broder, Marcus Fontoura and Vanja Josifovski. A Search-based Method for Forecasting Ad Impression in Contextual Advertising

Abstract: Contextual advertising} refers to the placement of small textual ads within the content of a generic web page. It has become a significant source of revenue for publishers ranging from individual bloggers to major newspapers. At the same time it is an important way for advertisers to reach their intended audience. This reach depends on the total number of exposures of the ad (impressions) and its click-through-rate (CTR) that can be viewed as the probability of an end-user clicking on the ad when shown. These two orthogonal, critical factors are both difficult to estimate and even individually can still be very informative in planning and budgeting advertising campaigns.

In this paper, we address the problem of forecasting the number of impressions for new or changed ads in the system. Producing such forecasts, even within large margins of error, is quite challenging:1) ad selection in contextual advertising is a complicated process based on tens or even hundreds of page and ad features; 2) the publishers' content and traffic vary over time; and 3) the scale of the problem is daunting: over a course of a week it involves billions of impressions, hundreds of millions of distinct pages, hundreds of millions of ads, and varying bids of other competing advertisers. We tackle these complexities by simulating the presence of a given ad with its associated bid over weeks of historical data. We obtain an impression estimate by counting how many times the ad would have been

displayed if it were in the system over that period of time. We estimate this count by an efficient two-level search algorithm over the distinct pages in the data set. Experimental results show that our approach can accurately forecast the expected number of impressions of contextual ads in real time. We also show how this method can be used in tools for bid selection and ad evaluation.

Olivier Chapelle and Ya Zhang. A Dynamic Bayesian Network Click Model for Web Search Ranking

Abstract: As with any application of machine learning, web search ranking requires labeled data. The labels usually come in the form of relevance assessments made by editors. Click logs can also provide an important source of implicit feedback and can be used as a cheap proxy for editorial labels. The main difficulty however comes from the so called position bias - urls appearing in lower positions are less likely to be clicked even if they are relevant.
In this paper, we propose a Dynamic Bayesian Network which aims at providing us with unbiased estimation of the relevance from the click logs. Experiments show that the proposed click model outperforms other existing click models in predicting both click-through rate and relevance.

hao yan, Shuai Ding and Torsten Suel. Inverted Index Compression and Query Processing with Optimized Document Ordering

Abstract: Web search engines use highly optimized compression schemes to decrease inverted index size and improve query throughput, and many index compression techniques have been studied in the literature. One approach taken by several recent studies [] first performs a renumbering of the document IDs in the collection that groups similar documents together, and then applies standard compression techniques. It is known that this can significantly improve index compression compared to a random document ordering.

We study index compression and query processing techniques for such reordered indexes. Previous work has focused on determining the best possible ordering of documents. In contrast, we assume that such an ordering is already given, and focus on how to optimize compression methods and query processing for this case. We perform an extensive study of compression techniques for document IDs. We also propose and evaluate techniques for compressing frequency values for this case. Finally, we study the effect of this approach on query processing performance. Our experiments show very significant improvements in index size and query processing speed on the TREC GOV2 collection of 25.2 million web pages.

QINGQING GAN and Torsten Suel. Improved Techniques for Result Caching in Web Search Engines

Abstract: Query processing is a major cost factor in operating large web search engines. In this paper, we study query result caching, one of the main techniques used to optimize query processing performance. Our first contribution is a study of result caching as a weighted caching problem in depth. Most previous work has focused on optimizing cache hit rates, but given that processing costs of queries can vary very significantly we argue that total cost savings also need to be considered. We describe and evaluate several algorithms for weighted result caching, and study the impact of Zipf-based query distributions on result caching. Our second and main contribution is a new set of feature-based cache eviction policies that achieve significant improvements over all previous methods, substantially narrowing the existing performance

gap to the theoretically optimal (clairvoyant) method. Finally, using the same approach, we also obtain performance gains for the related problem of inverted list caching.

Shuai Ding, Jinru He, Hao Yan and Torsten Suel. Using Graphics Processors for High Performance IR Query Processing

Abstract: Web search engines are facing formidable performance challenges due to data sizes and query loads. The major engines have to process tens of thousands of queries per second over tens of billions of documents. To deal with this heavy workload, such engines employ massively parallel systems consisting of thousands of machines. The significant cost of operating these systems has motivated a lot of recent research into more efficient query processing mechanisms. We investigate a new way to build such high performance IR systems using graphical processing units (GPUs). GPUs were originally designed to accelerate computer graphics applications through massive on-chip parallelism. Recently a number of researchers have studied how to use GPUs for other problem domains such as databases and scientific computing. Our contribution here is to design a basic system architecture for GPU-based high-performance IR, to develop suitable algorithms for subtasks such as inverted list compression, list intersection, and top-k scoring, and to show how to achieve highly efficient query processing on GPU-based systems. Our experimental results for a prototype GPU-based system on 25:2 million web pages indicate that significant gains in query processing performance can be obtained over a state-of-the-art CPU-based system.

Jay Chen, Lakshminarayanan Subramanian and Jinyang Li. RuralCafe: Web Search in the Rural Developing World

Abstract: A majority of the world's population in rural developing regions do not have access to the World Wide Web. Traditional network connectivity technologies have proven to be prohibitively expensive in these areas. The emergence of new long-range wireless technologies provide hope for connecting these rural regions to the Internet. However, the network connectivity provided by these new solutions are by nature {\em intermittent} due to high network usage rates, frequent power-cuts and the use of delay tolerant links.

Typical applications, especially interactive applications such as web search, cannot work in the face of intermittent connectivity. In this paper, we present the design and implementation of {\em RuralCafe}, a system intended to support efficient web search over intermittent networks. RuralCafe enables users to perform web search asynchronously and find what they are looking for in {\em one round of intermittency} as opposed to multiple rounds of search/downloads. RuralCafe does this by providing an expanded search query interface which allows a user to specify additional query terms to maximize the utility of the results returned by a search query. Given knowledge of the limited available network resources, RuralCafe performs optimizations to prefetch pages to best satisfy a search query based on a user's search preferences. In addition, RuralCafe does not require modifications to the web browser, and can provide single round search results tailored to various types of networks and economic constraints. We have implemented and evaluated the effectiveness of RuralCafe using queries from logs made to a large search engine, queries made by users in an intermittent setting, and live queries from a small testbed deployment.

Shengyue Ji, Guoliang Li, Chen Li and jianhua feng. Efficient Interactive Fuzzy Keyword Search

Abstract: Traditional information systems return answers after a user submits a

complete query. Users often feel ``left in the dark'' when they have limited knowledge about the underlying data, and have to use a try-and-see approach for finding information. A recent trend of supporting autocomplete in these systems is a first step towards solving this problem. In this paper, we study a new information-access paradigm, called ``interactive fuzzy search,'' in which the system searches the underlying data ``on the fly'' as the user types in query keywords. It extends autocomplete interfaces by (1) allowing keywords to appear in multiple attributes (in an arbitrary order) of the underlying data; and (2) finding relevant records that have keywords matching query keywords {\em approximately}. This framework allows users to explore data as they type, even in the presence of minor errors. We study research challenges in this framework for large amounts of data. Since each keystroke of the user could invoke a query on the backend, we need efficient algorithms to process each query within milliseconds. We develop various incremental-search algorithms using previously computed and cached results in order to achieve an interactive speed. We have deployed several real prototypes using these techniques. One of them has been deployed to support interactive search on the UC Irvine people directory, which has been used regularly and well received by users due to its friendly interface and high efficiency.

Sanjay Agrawal, Kaushik Chakrabarti, Surajit Chaudhuri, Venkatesh Ganti, Arnd Konig and Dong Xin. Exploiting Web Search Engines to Search Structured Information Sources

Abstract: Web search engines leverage information from structured databases to answer queries. For example, many product related queries on search engines ({\em Amazon, Google, Yahoo, Live Search}) are answered by searching underlying product databases containing names, descriptions, specifications, and reviews of products. However, these vertical search engines are ``{\em silo-ed}'' in that their results are independent of those from web search.
This often leads to empty or incomplete results, as query terms are matched against the information in the underlying database only.
In order to overcome this problem, we propose an approach that first identifies

relationships between web documents and items in structured databases. This allows us to subsequently exploit results from web search engines in combination with these relationships to obtain the structured data items relevant for a much wider range of queries. We propose an architecture that implements the integrated search functionality efficiently, adding very little additional overhead to query processing and is fully integrated with the search engine architecture. We demonstrate the quality of our techniques through an extensive experimental study.

Deepayan Chakrabarti, Ravi Kumar and Kunal Punera. Quicklink Selection for Navigational Query Results

Abstract: Quicklinks for a website are navigational shortcuts displayed below the website homepage on a search results page, and that let the users directly jump to selected points inside the website. Since the real-estate on a search results page is constrained and valuable, picking the best set of quicklinks to maximize the benefits for a majority of the users becomes an important problem for search engines. Using user browsing trails obtained from browser toolbars, and a simple probabilistic model, we formulate the quicklink selection problem as a combinatorial optimizaton problem. We first demonstrate the hardness of the objective, and then propose an algorithm that is provably within a factor of $(1-1/e)$ of the optimal. We also propose a different algorithm that works on trees and that can find the optimal solution; unlike the previous algorithm, this algorithm can incorporate natural constraints on the set of chosen quicklinks. The efficacy of our methods is demonstrated via empirical results on both a manually labeled set of websites and a set for which quicklink click through rates for several webpages were obtained from a real-world search engine.

Security and Privacy

Rich Gossweiler, Maryam Kamvar and Shumeet Baluja. What's Up CAPTCHA? A CAPTCHA Based on Image Orientation

Abstract: We present a new CAPTCHA where users adjust randomly rotated images to their upright orientation. Successfully determining the upright orientation requires the ability to analyze the complex contents of an image; a task that humans usually perform well and machines generally perform poorly. Given a large repository of images, such as those from a web search result, we use a suite of automated orientation detectors to prune those that can be easily set upright. We then apply a social feedback mechanism to verify that the remaining images have a humanrecognizable upright orientation. The main advantages of our CAPTCHA technique over the traditional text recognition technique are that it is language-independent, does not require text-entry (e.g. for a mobile device) and provides another domain for CAPTCHA generation beyond character obfuscation. This CAPTCHA lends itself to rapid implementation and has an almost limitless supply of images. We conducted extensive experiments regarding the viability of this technique and achieve positive results.

Leyla Bilge, Thorsten Strufe, Davide Balzarotti and Engin Kirda. All Your Contacts Are Belong to Us: Automated Identity Theft

Abstract: Social networking sites have been increasingly gaining popularity. Well-known sites such as Facebook have been reporting growth rates as high as 3% per week. Many social networking sites have millions of registered users who use these sites to share photographs, contact long-lost friends, establish new business contacts and to keep in touch. In this paper, we investigate how easy it would be for a potential attacker to launch automated crawling and identity theft (i.e., cloning) attacks against a number of popular social networking sites in order to gain access to a large volume of personal user information. The simplest attack we present consists of the automated identity theft of existing user profiles and the sending of friendship requests to the contacts of the cloned victim. The hope, from the attacker's point of view, is that the contacted users simply trust and accept the friendship request. By establishing a friendship relationship with the contacts of a victim, the attacker is able to access the sensitive personal information provided by them. In the second, more advanced attack we present, we show that it is effective and feasible to launch an automated, cross-site profile cloning attack. In this attack, we are able to automatically create a forged profile in a network where the victim is not registered yet and contact the victim's friends who are registered on both networks. Our experimental results with real users show that the automated attacks we present are effective and feasible in practice.

Balachander Krishnamurthy and Craig Wills. Privacy Diffusion on the Web: A Longitudinal Perspective

Abstract: For the last few years we have been studying the diffusion of privacy for users as they visit various Web sites triggering data gathering aggregation by third parties. This paper reports on our longitudinal study consisting of multiple snapshots of our examination of such diffusion over four years. We examine the various technical ways by which third-party aggregators acquire data and the depth of user-related information acquired. We study techniques for protecting privacy diffusion as well as limitations of such techniques. We

introduce the concept of secondary privacy damage.

Our results show increasing aggregation of user-related data by a steadily decreasing number of entities. A handful of companies are able to track users' movement across almost all of the popular Web sites. Virtually all the protection techniques have significant limitations highlighting the seriousness of the problem and the need for alternate solutions.

Arjun Guha, Shriram Krishnamurthi and Trevor Jim. Using Static Analysis for Ajax Intrusion Detection

Abstract: We present a static control-flow analysis for JavaScript programs running in a web browser. Our analysis tackles numerous challenges posed by modern web applications, including asynchronous communication, frameworks, and dynamic code generation. We use our analysis to extract a model of expected client behavior as seen from the server, and build an intrusion-prevention proxy for the server: the proxy intercepts client requests and disables those that do not meet the expected behavior. We insert random asynchronous requests to foil mimicry attacks. Finally, we evaluate our technique against several real applications and show that it protects against an attack in a widely-used web application.

Elena Zheleva and Lise Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles

Abstract: In order to address privacy concerns, many social media websites allow users to hide their personal profiles from the public. In this work, we show how an adversary can exploit an online social network with a mixture of public and private user profiles to predict the private attributes of users. We map this problem to a relational classification problem and we propose practical models that use friendship and group membership information (which is often not hidden) to infer sensitive attributes. The key novel idea is that in addition to friendship links, groups can be carriers of significant information. We show that on several well-known social media sites, we can easily and accurately recover the information of private-profile users. To the best of our knowledge, this is the first work that uses link-based and group-based classification to study privacy implications in social networks with mixed public and private user profiles.

anna squicciarini, Mohamed Shehab and Federica Paci. Collective Privacy Management in Social Networks

Abstract: Social Networking is one of the major technological phenomena of the Web 2.0, with hundreds of millions of people participating. Social networks enable a form of self expression for users, and help them to socialize and share content with other users.
In spite of the fact that content sharing represents one of the prominent features of existing Social Network sites, Social Networks yet do not support any mechanism for collaborative management of privacy settings for shared content.
In this paper, we model the problem of collaborative enforcement of privacy policies on shared data by using game theory. In particular, we propose a solution that offers automated ways to share images based on an extended notion of content ownership. Building upon the Clarke-Tax mechanism, we describe a simple mechanism that promotes truthfulness, and that rewards users who promote co-ownership.
We integrate our design with inference techniques that free the users from the burden of manually selecting privacy preferences for each picture.
To the best of our knowledge this is the first time such a protection mechanism for social networking has been proposed.
In the paper, we also show a proof-of-concept application, which we implemented

in the context of Facebook, one of today's most popular social networks. We show that supporting these type of solutions is not also feasible, but can be implemented through a minimal increase in overhead to end-users.

Guang Xiang and Jason Hong. A Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval

Abstract: Phishing is a significant security threat to the Internet, which causes tremendous economic loss every year. In this paper, we proposed a novel hybrid phish detection method based on state-of-the-art information extraction (IE) and information retrieval (IR) techniques. The identity-based component of our method detects phishing webpages by directly discovering the inconsistency between their identity and the identity they are imitating. The keywords-retrieval component utilizes IR algorithms exploiting the power of search engines to identify phish. Our method requires no training data, no prior knowledge of phishing signatures and specific implementations, and thus is able to adapt quickly to constantly appearing new phishing patterns. Comprehensive experiments over a diverse spectrum of data sources with 11449 pages show that both components have a low false positive rate and the stacked approach achieves a true positive rate of 90.06% with a false positive rate of 1.95%.

Semantic / Data Web

Sören Auer, Sebastian Dietzold, Jens Lehmann, Sebastian Hellmann and David Aumueller. Triplify - Light-weight Linked Data Publication from Relational Databases

Abstract: We present Triplify - a simplistic but effective approach to publish linked data from relational databases.
Triplify is based on mapping HTTP-URI requests onto relational database queries. Triplify transforms the resulting relations into RDF statements and publishes the data on the Web in various RDF serializations, in particular as Linked Data.
The rationale for developing Triplify is that the largest part of information on the Web is already stored in structured form, often as data contained in relational databases but published by Web applications merely as HTML mixing structure, layout and content.
In order to reveal the pure structured information behind the current Web we implemented Triplify as a light-weight software component, which can be easily integrated and deployed with the numerous widely installed Web applications. Our approach includes a method for publishing update logs to enable incremental crawling of linked data sources.
Triplify is complemented by a library of configurations for common relational schemata and a REST-enabled datasource registry.
Triplify configurations are provided containing mappings for many popular Web applications, including Wordpress, Drupal, Joomla, Gallery, and phpBB.
We show that despite its light-weight architecture Triplify is usable to publish very large datasets, such as 160GB of geo data from the OpenStreetMap project.

Senlin Liang, Paul Fodor, Hui Wan and Michael Kifer. OpenRuleBench: An Analysis of the Performance of Rule Engines

Abstract: The Semantic Web initiative has led to an upsurge of the interest in rules as a general and powerful way of processing, combining, and analyzing semantic information. Since several of the technologies underlying rule-based systems are already quite mature, it is important to understand how such systems might perform on the Web scale. OpenRuleBench is a suite of benchmarks for analyzing the performance and scalability of different rule engines. Currently the study spans five different technologies and eleven systems, but OpenRuleBench is an open community resource, and contributions from the community are welcome. In this paper, we describe the tested systems and technologies, the methodology used in testing, and analyze the results.

Fabian M. Suchanek, Mauro Sozio and Gerhard Weikum. SOFIE: Self-Organizing Flexible Information Extraction

Abstract: This paper presents SOFIE, a system that can extend an existing ontology by new facts.
SOFIE provides a integrative framework, in which information extraction, word disambiguation and semantic
reasoning all become part of one unifying model.
SOFIE processes text or Web sources and finds meaningful patterns. It maps the words in the pattern to entities in the ontology. It hypothesizes on the meaning
of the pattern, and checks the semantic plausibility of the hypothesis with the existing
ontology. Then the new fact is added to the ontology, avoiding inconsistency with the existing facts.
The logical model that connects existing facts, new hypotheses, extraction patterns, and consistency constraints is represented as a set of propositional clauses.

We use an approximation algorithm for the Weighted MAX SAT problem
to compute the most plausible subset of hypotheses.
Thereby, the SOFIE framework integrates the paradigms of pattern matching,
entity disambiguation, and ontological reasoning into one unified model,
and enables the automated growth of large ontologies.
Experiments, using the YAGO ontology as existing knowledge and
various text and Web corpora as input sources, show that our method
yields very good precision around 90 percent or higher.

Danh Le Phuoc, Axel Polleres, Christian Morbidoni, Manfred Hauswirth and Giovanni Tummarello.
Rapid Semantic Web Mashup Development through Semantic Web Pipes

Abstract: The use of RDF data published on the Web for applications is still a cumbersome and resource-intensive task due to the limited software support and the lack of standard programming paradigms to deal with everyday problems such as combination of RDF data from di erent sources, object identi er consolidation, ontology alignment and mediation or plain querying and processing tasks. While in a lot of other areas such tasks are supported by excellent libraries and component-oriented toolboxes of basic processing functionalities, RDF-basedWeb applications are still largely customized programs for a speci c purpose, with little potential for reuse. This increases development costs and incurs a more error-prone development process. Speaking in software engineering terms, this means that a good standard architectural style with good support for rapid application development is still missing. In this paper we present a framework based on the classical abstraction of pipes which tries to remedy this problem and support the fast implementation of software, while preserving desirable properties such as abstraction, encapsulation, component-orientation, code re-usability and maintainability, which are common and well supported in other application areas.

Maria Grineva, Dmitry Lizorkin and Maxim Grinev. Extracting Key Terms From Noisy and Multitheme Documents

Abstract: We present a novel method for key term extraction from text documents. In our method, document is modeled as a graph of semantic relationships between terms of that document. We exploit the following remarkable feature of the graph: the terms related to the main topics of the document tend to bunch up into densely interconnected subgraphs or communities, while non-important terms fall into weakly intercon-nected communities, or even become isolated vertices. We apply graph community detection techniques to partition the graph into thematically cohesive groups of terms. We introduce a criterion function to select groups that contain key terms discarding groups with unimportant terms. To weight terms and determine semantic relatedness between them we exploit information extracted from Wikipedia.
Using such an approach gives us the following two advantages. First, it allows effectively processing multi-theme documents. Second, it is good at filtering out noise information in the document, such as, for example, navigational bars or headers in web pages.
Evaluations of the method show that it outperforms existing methods producing key terms with higher precision and recall. Additional experiments on web pages prove that our method appears to be substantially more e ective on noisy and multi-theme documents than existing methods.

Jorge Gracia, Mathieu d'Aquin and Eduardo Mena. Large Scale Integration of Senses for the Semantic Web

Abstract: Nowadays, the increasing amount of semantic data available on the Web leads to a new stage in the potential of Semantic Web applications. However, it also introduces new issues due to the heterogeneity of the available semantic resources. One of the most remarkable is redundancy, that is, the excess of different semantic

descriptions, coming from different sources, to describe the same intended meaning.

In this paper, we propose a technique to perform a large scale integration of senses (expressed as ontology terms), in order to cluster the most similar ones, when indexing large amounts of online semantic information. It can dramatically reduce the redundancy problem on the current Semantic Web. In order to make this objective feasible, we have studied the adaptability and scalability of our previous work on sense integration, to be translated to the much larger scenario of the Semantic Web. Our evaluation shows a good behaviour of these techniques when used in large scale experiments, then making feasible the proposed approach.

Philippe Cudre-Mauroux, Parisa Haghani, Michael Jost, Karl Aberer and Hermann de Meer. idMesh: Graph-Based Disambiguation of Linked Data

Abstract: We tackle the problem of disambiguating entities on the Web. We propose a user-driven scheme where graphs of entities -- represented by globally identifiable declarative artifacts -- self-organize in a dynamic and probabilistic manner. Our solution has the following two desirable properties: i) it lets end-users freely define associations between arbitrary entities and ii) it probabilistically infers entity relationships based on uncertain links using constraint-satisfaction mechanisms. We outline the interface between our scheme and the current data Web, and show how higher-layer applications can take advantage of our approach to enhance search and update of information relating to online entities. We describe a decentralized infrastructure supporting efficient and scalable entity disambiguation and demonstrate the practicability of our approach in a deployment over several hundreds of machines.

Ben Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho and Gerd Stumme. Emergent Semantics of Social Tagging

Abstract: Social bookmarking systems are becoming increasingly important data sources for
bootstrapping and maintaining Semantic Web applications. Their emergent information structures have become known as folksonomies. A key question for harvesting semantics from these systems is how to extend and adapt traditional notions of similarity to folksonomies, and which measures are best suited for applications such as community detection, navigation support, semantic search, user profiling and ontology learning. Here we build an evaluation framework to compare various general folksonomy-based similarity measures, which are derived from several established information-theoretic, statistical, and practical measures.
Our framework deals generally and symmetrically with users, tags, and resources. For evaluation purposes we focus on similarity between tags and between resources and consider different methods to aggregate annotations across users. After comparing the ability of several tag similarity measures to predict
user-created tag relations, we provide an external grounding by user-validated semantic proxies based on WordNet and the Open Directory Project. We also investigate the issue of scalability. We find that mutual information with distributional micro-aggregation across users yields the highest accuracy, but is not scalable; per-user projection with collaborative aggregation provides the best scalable approach via incremental computations. The results are consistent across resource and tag similarity.

Social Networks and Web 2.0

Sharad Goel, Roby Muhamad and Duncan Watts. Social Search in "Small World" Experiments

Abstract: The "algorithmic small-world hypothesis" states that not only are pairs of individuals in a large social network connected by short paths, but that ordinary individuals can find these paths. Although theoretically plausible, empirical evidence for the hypothesis is limited, as most chains in "small-world" experiments fail to complete, thereby biasing estimates of "true" chain lengths. Using data from two recent small-world experiments, comprising a total of 162,328 message chains, and directed at one of 30 "targets" spread across 19 countries, we model heterogeneity in chain attrition rates as a function of individual attributes. We then introduce a rigorous way of estimating true chain lengths that is provably unbiased, and can account for empirically-observed variation in attrition rates. Our findings provide mixed support for the algorithmic hypothesis. On the one hand, it appears that roughly half of all chains can be completed in 6-7 steps--thus supporting the "six degrees of separation" assertion--but on the other hand, estimates of the mean are much longer, suggesting that for at least some of the population, the world is not "small" in the algorithmic sense. We conclude that search distances in social networks are fundamentally different from topological distances, for which the mean and median of the shortest path lengths between nodes tend to be similar.

Ulrik Brandes, Patrick Kenis, Juergen Lerner and Denise van Raaij. Network Analysis of Collaboration Structure in Wikipedia

Abstract: In this paper we give models and algorithms to describe and analyze the collaboration among authors of Wikipedia from a network analytical perspective. The edit-network encodes who interacts how with whom when editing an article; it significantly extends previous network-models that code author communities in Wikipedia. Several characteristics summarizing some aspects of the organization process and allowing the analyst to identify certain types of authors can be obtained from the edit-network. Moreover, we propose several indicators characterizing the global network structure and methods to visualize edit-networks. It is shown that the structural network indicators are correlated with quality labels of the associated Wikipedia articles.

Yutaka Matsuo and Hikaru Yamamoto. Community Gravity: Measuring Bidirectional Effects by Trust and Rating on Online

Abstract: Several attempts have been made to analyze customer behavior on online E-commerce sites. Some studies particularly emphasize the social networks of customers. Users' reviews and ratings of a product exert effects on other consumers' purchasing behavior. Whether a user refers to other users' ratings depends on the trust accorded by a user to the reviewer. On the other hand, the trust that is felt by a user for another user correlates to the similarity of ratings by two users. This bidirectional interaction that involves trust and rating is an important aspect of understanding consumer behavior in online communities because it suggests clustering of similar users and the evolution of strong communities. This paper presents a theoretical model along with analyses of an actual online E-commerce site. We analyzed a large community site in Japan: @cosme. The noteworthy characteristics of @cosme are that users can bookmark their trusted users; in addition, they can post their ratings of products, which facilitates our analyses of the ratings' bidirectional effects on trust and ratings. We describe an overview of the data in @cosme, analyses of effects from trust to rating and vice versa, and our proposition of a measure of of community gravity, which measures how strongly a user might be attracted to a community. Our study is based on the @cosme dataset in addition to the Epinions dataset. It elucidates

important insights and proposes a potentially important measure for mining online social networks.

Shilad Sen, Jesse Vig and John Riedl. Tagommenders: Connecting Users to Items through Tags

Abstract: Tags have emerged as a powerful mechanism that enable users to find, organize, and understand online entities. Recommender systems similarly enable users to efficiently navigate vast collections of items. Algorithms combining tags with recommenders may deliver both the automation inherent in recommenders, and the flexibility and conceptual comprehensibility inherent in tagging systems. In this paper we explore tagommenders, recommender algorithms that predict users' preferences for items based on their inferred preferences for tags. We describe tag preference inference algorithms based on users' interactions with tags and movies, and evaluate these algorithms based on tag preference ratings collected from 995 MovieLens users. We design and evaluate algorithms that predict users' ratings for movies based on their inferred tag preferences. Our research promises to increase the accuracy of recommendations in tagging systems, and it may lead to flexible recommender systems that leverage the characteristics of items users find most important.

Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein and Hongyuan Zha. Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement

Abstract: Community Question Answering (CQA) has emerged as a popular forum for users to pose questions for other users to answer. Over the last few years, CQA portals such as Naver and Yahoo! Answers have exploded in popularity, and now provide a viable alternative to general purpose web search. At the same time, the answers to past questions submitted in CQA sites comprise a valuable knowledge repository which could be a gold mine for information retrieval and automatic question answering. Unfortunately, the quality of the submitted questions and answers varies widely - increasingly so that a large fraction of the content is not usable for answering queries. Previous approaches for retrieving relevant and high quality content have been proposed, but they require large amounts of manually labeled data -- which limits the applicability of the supervised approaches to new sites and domains. In this paper we address this problem by developing a semi-supervised coupled mutual reinforcement framework for simultaneously calculating content quality and user reputation, that requires relatively few labeled examples to initialize the training process. Results of a large scale evaluation demonstrate that our methods are more effective than previous approaches for finding high-quality answers, questions, and users. More importantly, our quality estimation significantly improves the accuracy of search over CQA archives over the state-of-the-art methods.

Jérôme Kunegis, Andreas Lommatzsch and Christian Bauckhage. The Slashdot Zoo: Mining a Social Network with Negative Edges

Abstract: We analyse the corpus of user relationships of the Slashdot technology news site.
The data was collected from the Slashdot Zoo feature
where users of the website can tag other users as friends and foes,
providing positive and negative endorsements.
We adapt social network analysis techniques to the problem of negative edge weights.
In particular, we consider signed variants
of global network characteristics such as the clustering coefficient,
node-level characteristics such as centrality and popularity measures,
and link-level characteristics such as distances and similarity measures.

We evaluate these measures on the task of identifying unpopular users, as well as on the task of predicting the sign of links and show that the network exhibits multiplicative transitivity which allow algebraic methods based on matrix multiplication to be used.
We compare our methods to traditional methods which are only suitable for positively weighted edges.

Dietwig Lowet and Daniel Goergen. Co-browsing dynamic web pages

Abstract: Collaborative browsing, or co-browsing, is the co-navigation of the web with other people at-a-distance, supported by software that takes care of synchronizing the browsers. Current state-of-the-art solutions are able to do co-browsing of "static web pages", and do not support the synchronization of JavaScript interactions. Currently many web pages use JavaScript and Ajax techniques to create highly dynamic web applications. In this paper, we describe two approaches for co-browsing that both support the synchronization of the JavaScript and Ajax interactions of dynamic web pages. One approach is based on synchronizing the output of the JavaScript engine by sending over the changes made on the DOM tree. The other approach is based on synchronizing the input of the JavaScript engine by synchronizing UI events and incoming data. Since the latter solution offers a better user experience and is more scalable, it is elaborated in more detail. An important aspect of both approaches is that they operate at the DOM level. Therefore, the client-side can be implemented in JavaScript and no browser extensions are required. To the best of the authors' knowledge this is the first DOM-level co-browsing solution that also enables co-browsing of the dynamic interaction parts of web pages. The presented co-browsing solution has been implemented in a research demonstrator which allows users to do co-browsing of web-applications on browser-based networked televisions

Anon Plangprasopchok and Kristina Lerman. Constructing Folksonomies from User-specified Relations on Flickr

Abstract: Automatic folksonomy construction from tags has attracted much attention recently. However, inferring hierarchical relations between concepts from tags has a drawback in that it is difficult to distinguish between more popular and more general concepts. Instead of tags we propose to use user-specified relations for learning folksonomy.
We explore two statistical frameworks for aggregating many shallow individual hierarchies, expressed through the collection/set relations on the social photosharing site Flickr, into a common deeper folksonomy that reflects how a community organizes knowledge. Our approach addresses a number of challenges that arise while aggregating information from diverse users, namely noisy vocabulary, and variations in the granularity level of the concepts expressed.
Our second contribution is a method for automatically evaluating learned folksonomy by comparing it to a reference taxonomy, e.g., the Web directory created by the Open Directory Project. Our empirical results suggest that user-specified relations are a good source of evidence for learning folksonomies.

Jose San Pedro and Stefan Siersdorfer. Ranking and Classifying Attractiveness of Photos in Folksonomies

Abstract: Web 2.0 applications like Flickr, YouTube, or Del.icio.us are increasingly popular online communities for creating, editing and sharing content and the growing size of these folksonomies poses new challenges in terms of search and mining. In this paper we introduce a novel methodology for automatically ranking and classifying photos according to their attractiveness for folksonomy members. To this end, we exploit image features known for having significant effects on the visual quality perceived by humans (e.g. sharpness and colorfulness) as well as textual meta data, in what is a multi-modal approach. Using feedback and annotations available in the Web 2.0 photo sharing system Flickr, we assign

relevance values to the photos and train classification and regression models based on these relevance assignments. With the resulting machine learning models we categorize and rank photos according to their attractiveness. Applications include enhanced ranking functions for search and recommender methods for attractive content. Large scale experiments on a collection of Flickr photos demonstrate the viability of our approach.

David Crandall, Lars Backstrom, Daniel Huttenlocher and Jon Kleinberg. Mapping the World's Photos

Abstract: We investigate how to organize a large collection of geotagged photos, working with a dataset of about 20 million images collected from Flickr. Our approach combines content analysis based on text tags and image data with structural analysis based on geospatial data. We use the spatial distribution of where people take photos to define a relational structure between the photos that are taken at popular places. We then study the interplay between this structure and the content, using classification methods for predicting such locations from visual, textual and temporal features of the photos. We find that visual and temporal features improve the ability to estimate the location of a photo, compared to using just textual features. We illustrate using these techniques to organize a large photo collection, while also revealing various interesting properties about popular cities and landmarks at a global scale.

Munmun De Choudhury, Hari Sundaram, Ajita John and Doree Seligmann. What Makes Conversations Interesting? Themes, Participants and Consequences of Conversations in Online Social Media

Abstract: Rich media social networks promote not only creation and consumption of media, but also communication about the posted media item. What causes a conversation to be interesting, that prompts a user to participate in the discussion on a posted video? We conjecture that people will participate in conversations when they find the conversation theme interesting, see comments by people that are known to them or observe an engaging dialogue between two or more people is engaging (an absorbing back and forth between two people). Importantly, a conversation that is deemed interesting must be consequential – i.e. it must impact the social network itself.
Our framework has three parts: characterizing themes, characterizing participants for determining interestingness and measures of consequences of a conversation deemed to be interesting. First, we detect conversational themes using a sophisticated mixture model approach. Second, we determine interestingness of participants and interestingness of conversations based on a random walk model. Third, we measure the consequence of a conversation by measuring the mutual information of the interesting property with three variables that should be affected by an interesting conversation – participation in related themes, participant cohesiveness and theme diffusion. We have conducted extensive experiments using dataset from the popular video sharing site, YouTube. Our results show that our method of interestingness maximizes the mutual information, and is significantly better (twice as large) than three other baseline methods (number of comments, number of new participants and PageRank based assessment).

Thomas Karagiannis and Milan Vojnovic. Behavioral Profiles for Advanced Email Features

Abstract: We examine the behavioral patterns of email usage in a large-scale enterprise over a three-month period. In particular, we focus on two main questions: (Q1) what do replies depend on? and (Q2) what is the gain of augmenting contacts through the friends of friends from the email social

graph? For Q1, we identify and evaluate the significance of several factors that affect the reply probability and the email response time. We find that all factors of our considered set are significant, provide their relative ordering, and identify the recipient list size, and the intensity of email communication between the correspondents as the dominant factors. We highlight various novel threshold behaviors and provide support for existing hypotheses such as that of the least-effort reply. For Q2, we find that the number of new contacts extracted from the friends-of-friends relationships amounts to a large number, but which is still a limited portion of the total enterprise size. We believe that our results provide significant insights towards informed design of advanced email features, including those of social-networking type.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg and Lillian Lee. How opinions are received by online communities: A case study on Amazon.com helpfulness votes

Abstract: There are many on-line settings in which users publicly express opinions. A number of these offer mechanisms for other users to evaluate these opinions; a canonical example is Amazon.com, where reviews come with annotations like ``26 of 32 people found the following review helpful.'' Opinion evaluation appears in many off-line settings as well, including market research and political campaigns. Reasoning about the evaluation of an opinion is fundamentally different from reasoning about the opinion itself: rather than asking, ``What did Y think of X?'', we are asking ``What did Z think of Y's opinion of X?''

Here we develop a framework for analyzing and modeling opinion evaluation, using a large-scale collection of Amazon book reviews as a dataset. We find that the perceived helpfulness of a review is based not just on its content but also depends in subtle ways on its score relative to other scores for the same product. As part of our approach, we develop novel methods to control for the effects of text in opinion evaluation, and we provide a simple, natural mathematical model consistent with our findings. Our analysis also allows us to distinguish among the predictions of competing theories from sociology and social psychology, and to discover unexpected differences in the collective opinion evaluation behavior of user populations from different countries.

Meeyoung Cha, Alan Mislove and Krishna Gummadi. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network

Abstract: Online social networking sites like MySpace, Facebook and Flickr have become a popular way to share and disseminate content. Their massive popularity has led to viral marketing techniques that attempt to spread content, products, and political campaigns on these sites. However, there is little data publicly available on viral propagation in the real world and few studies have attempted to characterize how information spreads over current online social networks.
In this paper, we collect and analyze large-scale traces of information dissemination in the Flickr social network. Our analysis, based on repeated crawls of the favorite markings of 11 million photos and the social network of 2.5 million users, attempts to answer three key questions: (a) how widely does information propagates in the social network? (b) how quickly does information propagate? and (c) what is the role of word-of-mouth exchanges between friends in the overall propagation of information in the network? Somewhat counter intuitively, we find that (a) even popular information does not spread widely throughout the network, (b) even popular photos spread very slowly through the network, and (c) information exchanged between friends is likely to account for over 50\% of all favorite

bookmarks. Our findings stand in sharp contradiction to our initial expectation that information would spread widely and quickly in a viral fashion across the social network.

Web Engineering

Woralak Kongdenfha, Boualem Benatallah, Julien Vayssière, Regis Saint-Paul and Fabio casati. Rapid Development of Spreadsheet-based Web Mashups

Abstract: The rapid growth of social networking sites and web communities have motivated web sites to expose their APIs to external developers who create mashups by assembling existing functionalities. Current APIs, however, aim toward developers with programming expertise; they are not directly usable by wider class of users who do not have programming background, but would nevertheless like to build their own mashups. To address this need, we propose a spreadsheet-based Web mashups development framework, which enables users to develop mashups in the popular spreadsheet environment. Our contributions are as follows. First, we provide a mechanism that makes structured data first class value of spreadsheet cells. Second, we propose a new component model that can be used to develop fairly sophisticated mashups, involving joining data sources, and keeping data presenting on spreadsheets up to date. Third, to simplify mashup development, we provide a collection of spreadsheet-based mashup patterns that capture common Web data access and spreadsheet presentation functionalities. Users can reuse and customize these patterns to build spreadsheet-based Web mashups instead of developing them from scratch. Fourth, we enable users to manipulate structured data presented on spreadsheet in a drag-and-drop fashion. Finally, we have developed and tested a proof-of-concept prototype to demonstrate the utility of the proposed framework.

Cesare Pautasso and Erik Wilde. Why is the Web Loosely Coupled? A Multi-Faceted Metric for Service Design

Abstract: Loose coupling is often quoted as a desirable property of systems architectures. One of the main goals of building systems using Web technologies is to achieve loose coupling. However, given the lack of a widely accepted definition of this term, it becomes hard to use coupling as a criterion to evaluate alternative Web technology choices, as all options may exhibit, and claim to provide, some kind of "loose" coupling effects. This paper presents a systematic study of the degree of coupling found in service-oriented systems based on a multi-faceted approach. Thanks to the metric introduced in this paper, coupling is no longer a one-dimensional concept with loose coupling found somewhere in between tight coupling and no coupling. The paper shows how the metric can be applied to real-world examples in order to support and improve the design process of service-oriented systems.

Mohammad Alrifai and Thomas Risse. Combining Global Optimization with Local Selection for Efficient QoS-aware Service Composition

Abstract: The run-time binding of web services has been recently put forward in order to support rapid and dynamic web service compositions. With the growing number of alternative web services that provide the same functionality but differ in quality parameters, the service composition becomes a decision problem on which component services should be selected such that user's end-to-end QoS requirements (e.g. availability, response time) and preferences (e.g. price) are satisfied. Although very efficient, local selection strategy fails short in handling global QoS requirements. Solutions based on global optimization, on the other hand, can handle global constraints, but their poor performance renders them inappropriate for applications with dynamic and real-time requirements. In this

paper we address this problem and propose a solution that combines global optimization with local selection techniques to benefit from the advantages of both worlds. The proposed solution consists of two steps: first, we use mixed integer programming (MIP) to find the optimal decomposition of global QoS constraints into local constraints. Second, we use distributed local search to find the best web services that satisfy these local constraints. The results of experimental evaluation indicate that our approach significantly outperforms existing solutions in terms of computation time while achieving close-to-optimal results.

Guiling Wang, Shaohua Yang and Yanbo Han. Mashroom: End-User Mashup Programming Using Nested Tables

Abstract: This paper presents an end-user-oriented programming environment called Mashroom. Major contributions herein include an end-user programming model with an expressive and graphical data structure as well as a set of formally-defined visual mashup operators. The data structure takes advantage of nested table, and maintains the intuitiveness while allowing users to express complex data objects. The mashup operators visualized with contextual menu and formula bar are directly applied on the data. Experiments and case studies reveal that end users have little difficulty in effectively and efficiently using Mashroom to build mashup applications.

Jalal Mahmud, Yevgen Borodin, I.V. Ramakrishnan and C. R. Ramakrishnan. Automated Construction of Web Accessibility Models from Transaction Click Streams

Abstract: Screen readers, the dominant assistive technology used by visually impaired users to access theWeb, function by speaking out the content of the screen serially. Using screen readers for conducting online transactions can cause considerable information overload, because transactions, such as shopping and paying bills, typically involve a number of steps spanning several web pages. One can combat this overload with a transaction model for web accessibility that presents only fragments of web pages that are needed for doing transactions. We can realize such a model by coupling a process automata, encoding states of a transaction, with concept classifiers that identify page fragments "relevant" to a particular state of the transaction.

In this paper we present a fully automated process that synergistically combines several techniques for transforming click stream data generated by transactions into a transaction model. These techniques include web content analysis to partition a web page into segments consisting of semantically related content elements, contextual analysis of data surrounding clickable objects in a page, and machine learning methods, such as clustering of page segments based on contextual analysis, statistical classification, and automata learning.

A unique aspect of the transformation process is that the click streams, that serve as the training data for the learning methods, need not be labeled. More generally, it operates with partially labeled click stream data where some or all the labels could be missing. Not having to rely exclusively on (manually) labeled click stream data has important benefits: (i) visually impaired users do not have to depend on sighted users for the training data needed to construct transaction models; (ii) it is possible to mine personalized models from transaction click streams associated with sites that visually impaired users visit regularly; (iii) since partially labeled data is relatively easy to obtain, it is feasible to scale up the construction of domain-specific transaction models (e.g.: separate models for shopping, airline reservations, bill payments, etc.); (iv)

adjusting the performance of deployed models over time with new training data is also doable.
We provide preliminary experimental evidence of the practical effectiveness of both domain-specific, as well as personalized accessibility transaction models built using our approach. Finally, this approach is applicable for building transaction models for mobile devices with limited-size displays, as well as for creating wrappers for information extraction from web sites.

Ghislain Fourny, Markus Pilman, Daniela Florescu, Donald Kossmann, Tim Kraska and Darin McBeath. XQuery in the Browser

Abstract: Since the invention of the Web, the browser has become more and more powerful. By now, it is a programming and execution environment in itself. The predominant language to program applications in the browser today is JavaScript. With browsers becoming more powerful, JavaScript has been extended and new layers have been added (e.g., DOM-Support and XPath). Today, JavaScript is very successful and applications and GUI features implemented in the browser have become increasingly complex. The purpose of this paper is to improve the programmability of Web browsers by enabling the execution of XQuery programs in the browser. Although it has the potential to ideally replace JavaScript, it is possible to run it in addition to JavaScript for more flexibility. Furthermore, it allows instant code migration from the server to the client and vice-versa. This enables a significant simplification of the technology stack. The intuition is that programming the browser involves mostly XML (i.e., DOM) navigation and manipulation, and the XQuery family of W3C standards were designed exactly for that purpose. The paper proposes extensions to XQuery for Web browsers and gives a number of examples that demonstrate the usefulness of XQuery for the development of AJAX-style applications. Furthermore, the paper presents the design of an XQuery plug-in for Microsoft's Internet Explorer. The paper also gives examples of applications which were developed with the help of this plug-in.

Michiaki Tatsubori and Toyotaro Suzumura. HTML Templates that Fly - A Template Engine Approach to Automated Offloading from Server to Client

Abstract: Web applications often use HTML template engines to separate the webpage presentation from its underlying business logic and objects. This is now the de facto standard programming model for Web application development. This paper proposes a novel implementation for existing server-side template engines, MoveableTemplate, for (a) reduced bandwidth consumption in Web application servers, and (b) off-loading HTML generation tasks to Web clients. Instead of producing a fully-generated HTML page, the proposed template engine produces a skeletal script which includes only the dynamic values of the template parameters and the bootstrap code that runs on a Web browser at client side. It retrieves a client-side template engine and the payload templates separately. With the goals of efficiency, implementation transparency, security, and standard compliance in mind, we developed MoveableTemplate with two design principles: effective browser cache usage, and reasonable compromises which restrict template usage patterns and relax security policies slightly but explicitly. This approach allows most typical template-based Web applications to run effectively with MoveableTemplate. As an experiment, we tested the SPECweb2005 Banking application using MoveableTemplate without any other modifications and saw throughput improvements from 1.5x to 2.0x at its best mode. Moreover, MoveableTemplate makes applications comply with a specified security policy while conventional technologies for automatic client-server partitioning can pose security problems in the Web environment.

William Conner, Arun Iyengar, Thomas Mikalsen, Isabelle Rouvellou and Klara Nahrstedt. A Trust Management Framework for Service-Oriented Environments

Abstract: Many reputation management systems have been developed under the assumption that each entity in the system will use a variant of the same scoring function. Much of the previous work in reputation management has focused on providing robustness and improving performance for a given reputation scheme. In this paper, we present a reputation-based trust management framework that supports the synthesis of trust-related feedback from many different entities while also providing each entity with the flexibility to apply different scoring functions over the same feedback data for customized trust evaluations. We also propose a novel scheme to cache trust values based on recent client activity. To evaluate our approach, we implemented our trust management service and tested it on a realistic application scenario in both LAN and WAN distributed environments. Our results indicate that our trust management service can effectively support multiple scoring functions with low overhead and high availability.

Chuan Yue and Haining Wang. Characterizing Insecure JavaScript Practices on the Web

Abstract: JavaScript is an interpreted programming language most often used for enhancing web page interactivity and functionality. It has powerful capabilities to interact with web page documents and browser windows, however, it has also opened the door for many browser-based security attacks. Insecure engineering practices of using JavaScript may not directly lead to security breaches, but can create new attack vectors and greatly increase the risks of browser-based attacks. In this paper, we present the first measurement study on insecure practices of using JavaScript on the web. Our focus is on the insecure practices of JavaScript inclusion and dynamic generation, and we examine their severity and nature on 6,805 unique web sites. Our measurement results reveal that insecure JavaScript practices are common at various web sites: (1) at least 66.4% of the measured web sites manifest the insecure practices of including JavaScript files from external domains into the top-level documents of their web pages; (2) over 44.4% of the measured web sites use the dangerous eval() function to dynamically generate and execute JavaScript code on their web pages; and (3) in JavaScript dynamic generation, using the document.write() method and innerHTML property is much more popular than using the relatively secure technique of creating script elements via DOM methods. Our analysis indicates that safe alternatives to these insecure practices exist in common cases and ought to be adopted by web site developers and administrators for reducing potential security risks.

Heiko Ludwig, Jim Laredo, Kamal Bhattacharya, Liliana Pasquale and Bruno Wassermann. REST-Based Management of Loosely Coupled Services

Abstract: Applications increasingly make use of the distributed platform that the World Wide Web provides – be it as a Software-as-a-Service such as salesforce.com, an application infrastructure such as facebook.com, or a computing infrastructure such as a "cloud". A common characteristic of applications of this kind is that they are deployed on infrastructure or make use of components that reside in different management domains. Current service management approaches and systems, however, often rely on a centrally managed configuration management database (CMDB), which is the basis for centrally orchestrated service management processes, in particular change management and incident management. The distribution of management responsibility of WWW based applications requires a decentralized approach to service management. This paper proposes an approach of decentralized service management based on distributed configuration management and service process co-ordination, making use RESTful access to configuration information and ATOM-based distribution of updates as a novel foundation for service management processes.

Lijun Mei, Zhenyu Zhang, W.K. Chan and T.H. Tse. Test Case Prioritization in Regression Testing of Service-Oriented Business Applications

Abstract: Regression testing assures the quality of modified service-oriented business applications against unintended changes. However, a typical regression test suite is large in size. Executing those test cases that may detect failures earlier is attractive. Many existing prioritization techniques arrange test cases based on their individual code coverage on program statements of a previous version of the application. On the other hand, industrial service-oriented business applications are typically written in orchestration languages such as WS-BPEL and integrate workflow steps and web services via XPath and WSDL. Faults in these artifacts may cause the application to extract wrong data from messages and lead to failures in executing service compositions. Surprisingly, little existing regression testing research considers these artifacts. We propose a multilevel coverage model to capture business process, XPath and WSDL from the regression testing perspective. Atop the model, we develop a family of test case prioritization techniques. The empirical results show that our techniques can achieve significantly higher rates of fault detection than existing techniques.

XML and Web Data

Huayu Wu, TokWang Ling, Liang Xu and Zhifeng Bao. Performing grouping and aggregate functions in XML queries

Abstract: Since more and more business data are represented in XML format, there is a compelling need of supporting analytical operations in XML queries. Particularly, the latest version of XQuery proposed by W3C, XQuery 1.1, introduces a new construct to explicitly express grouping operation in FLWOR expression. Existing works in XML query processing mainly focus on physically matching query structure over XML document. Given the explicit grouping operation in a query, how to efficiently compute grouping and aggregate functions over XML document is not well studied yet. In this paper, we extend our previous XML query processing algorithm, VERT, to efficiently perform grouping and aggregate function in queries. The main technique of our approach is introducing relational tables to index values. Query pattern matching and aggregation computing are both conducted with table indexes. We also propose two semantic optimizations to further improve the query performance. Finally we present experimental results to validate the efficiency of our approach, over other existing approaches.

Gengxin Miao, Junichi Tatemura, Wang-Pin Hsiung, Arsany Sawires and Louise Moser. Extracting Data Records from the Web Using Tag Path Clustering

Abstract: Fully automatic methods to extract lists of objects from the Web have been studied extensively. Record extraction, the first step of this object extraction process, identifies a set of page segments, each of which represents an individual object (e.g., a product). State-of-the-art methods suffice for simple search, but they often fail to handle more complicated or noisy page structures due to a key limitation -- their greedy manner of identifying a list of records through pairwise comparison (i.e., similarity match) of consecutive segments. This paper introduces a new method for record extraction that captures a list of objects in a more robust way based on a holistic analysis of a Web page. The method focuses on how a distinct tag path appears repeatedly in the document DOM tree. Instead of comparing a pair of individual segments, it compares a pair of tag path occurrence patterns (called visual signals) to estimate how likely these two tag paths represent the same list of objects. The paper introduces a similarity measure that captures how closely the visual signals appear and interleave. Clustering of tag paths is then performed based on this similarity measure, and sets of tag paths that form the structure of data records are extracted. Experiments show that this method achieves higher accuracy than previous work.

Uri Schonfeld and Narayanan Shivakumar. Sitemaps: Above and Beyond the Crawl of Duty

Abstract: Comprehensive coverage of the public web is crucial to web search engines. Search engines use crawlers to retrieve pages and then discover new ones by extracting the pages' outgoing links. However, the set of pages reachable from the publicly linked web is estimated to be significantly smaller than the invisible web, the set of documents that have no incoming links and can only be retrieved through web applications and web forms. The Sitemaps protocol is a fast-growing web protocol supported jointly by major search engines to help content creators and search engines unlock this hidden data by making it available to search engines. In this paper, we perform a detailed study of how "classic" discovery crawling compares with Sitemaps, in key measures such as coverage and freshness over key representative websites as well as over billions of URLs seen at Google. We observe that Sitemaps and discovery crawling complement each other very well, and offer different tradeoffs.

Jeff Pasternack and Dan Roth. Extracting Article Text from the Web with Maximum

Subsequence Segmentation

Abstract: Much of the information on the Web is found in articles from online news outlets, magazines, encyclopedias, review collections, and other sources. However, extracting this content from the original HTML document is complicated by the large amount of less informative and typically unrelated material such as navigation menus, forms, user comments, and ads. Existing approaches tend to be either brittle and demand significant expert knowledge and time (manual or tool-assisted generation of rules or code), necessitate labeled examples for every different page structure to be processed (wrapper induction), require relatively uniform layout (template detection), or, as with Visual Page Segmentation (VIPS), are computationally expensive. We introduce maximum subsequence segmentation, a method of global optimization over token-level local classifiers, and apply it to the domain of news websites. Training examples are easy to obtain, both learning and prediction are linear time, and results are excellent (our semi-supervised algorithm yields an overall F1-score of 97.947%), surpassing even those produced by VIPS with a hypothetical perfect block-selection heuristic. We also evaluate against the recent CleanEval shared task with surprisingly good cross-task performance cleaning general web pages, exceeding the top "text-only" score (based on Levenshtein distance), 87.8% versus 84.1%.

Industrial Track - Practice and Experience

Wen-Yen Chen, Jon-Chyuan Chu, Junyi Luan, Hongjie Bai and Edward Chang. Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior

Abstract: Users of social networking services can connect with each other by forming communities for online interaction. Yet as the number of communities hosted by such websites grows over time, users have even greater need for effective community recommendations in order to meet more users. In this paper, we investigate two algorithms from very different domains and evaluate their effectiveness for personalized community recommendation. First is Association Rule Mining (ARM), which discovers associations between sets of communities that are shared across many users. Second is Latent Dirichlet Allocation (LDA), which models user-community co-occurrences using latent aspects. In comparing LDA with ARM, we are interested in discovering whether modeling latent structure is more effective for recommendations versus directly mining rules from the observed data. We experiment on an Orkut data set consisting of 492,104 users and 118,002 communities. We show that LDA consistently performs better than ARM using the top-k recommendations ranking metric, and we analyze examples of the latent information learned by LDA to explain this finding. To efficiently handle the large-scale data set, we parallelize LDA on distributed computers and demonstrate our parallel implementation's scalability with varying numbers of machines.

Murat Ali Bayir, ismail toroslu, Ahmet Cosar and Guven Fidan. SmartMiner: A New Framework for Mining Large Scale Web Usage Data

Abstract: In this paper, we propose a novel framework called Smart-Miner for web usage mining problem which uses link information for producing accurate user sessions and frequent navigation patterns. Unlike the simple session concepts in the time and navigation based approaches, where sessions are sequences of web pages requested from the server or viewed in the browser, in Smart-Miner sessions are set of paths traversed in the web graph that corresponds to users' navigations among web pages. We have modeled session reconstruction as a new graph problem and utilized a new algorithm, Smart-SRA, to solve this problem efficiently. For the pattern discovery phase, we have developed an efficient version of the Apriori-All technique which uses the structure of web graph to increase the performance. From the experiments that we have performed on both real and simulated data, we have observed that Smart-Miner produces at least 30%more accurateweb usage patterns than other approaches including previous session construction methods. We have also studied the effect of having the referrer information in the web server logs to show that different versions of Smart-SRA produce similar results. Another novel work is that we have implemented distributed version of the Smart Miner framework by employing Map-Reduce paradigm which enables processing huge size web server logs belonging to multiple web sites. To the best of our knowledge this paper is the first attempt to propose such large scale framework forweb usage mining problem. We conclude that we can efficiently process terabytes of web server logs belonging to multiple web sites by employing our scalable framework.

Wei Chu and Seung-Taek Park. Personalized Recommendation on Dynamic Contents Using Predictive Bilinear Models

Abstract: In Web-based services of dynamic contents (such as news articles), recommender systems always face the difficulties of timely identifying new items of high-quality and providing recommendations for new users. We propose a feature-based machine learning approach to personalized recommendation that is capable of handling the cold-start issues effectively. We maintain profiles of contents, in which temporal characteristics of contents, e.g. popularity or freshness, are updated in real-time manner. We also maintain profiles of users including demographical information and a summary of user activities within Yahoo! properties. Based on all features in user and content profiles, we develop predictive bilinear regression models to provide accurate personalized recommendations of new items for both existing and new users. This approach results in an offline model with light computational overhead compared with other recommender systems that require online re-training. The proposed framework is general and flexible for other personalized tasks. The superior performance of our approach is verified on a large-scale data set collected from the Today Module on Yahoo! Front Page, with comparison against six competitive approaches.

Web in IberoAmerica

Adriano Pereira, Diego Duarte, Wagner Meira Jr., Virgilio Almeida and Paulo Goes. Analyzing Seller Practices in a Brazilian Marketplace

Abstract: E-commerce is growing at an exponential rate. In the last decade, there has been an explosion of online commercial activity enabled by World Wide Web (WWW). These days, many consumers are less attracted to online auctions, preferring to buy merchandise quickly using fixed-price negotiations. Sales at Amazon.com, the leader in online sales of fixed-price goods, rose 37% in the first quarter of 2008. At eBay, where auctions make up 58% of the site's sales, revenue rose 14%. In Brazil, probably by cultural influence, online auctions are not been popular. This work presents a characterization and analysis of fixed-price online negotiations. Using actual data from a Brazilian marketplace, we analyze seller practices, considering seller profiles and strategies. We show that different sellers adopt strategies according to their interests, abilities and experience. Moreover, we confirm that choosing a selling strategy is not simple, since it is important to consider the seller's characteristics to evaluate the applicability of a strategy. The work also provides a comparative analysis of some selling practices in Brazil with popular worldwide marketplaces.

Guilherme Vale Menezes, Nivio Ziviani, Alberto Laender and Virgilio Almeida. A Geographical Analysis of Knowledge Production in Computer Science

Abstract: The goal of this paper is to analyze coauthorship networks in the Computer Science community. For this, we considered 30 graduate programs in Computer Science in different regions of the world, being 8 programs in Brazil, 16 in North America (3 in Canada and 13 in the United States) and 6 in Europe (2 in France, 1 in Switzerland and 3 in the United Kingdom). We collected data about 176,537 authors and 352,766 publication entries distributed among 2,176 publication venues. The results obtained from different measurements over the social networks show differences in the publication profile of Brazilian, European and North-American programs. For instance, the size of the giant component indicates the existence of isolated collaboration groups inside European programs, contrasting with their Brazilian and North-American counterparts. We also analyzed the temporal evolution of the social networks representing the three different regions. We observed that the number of collaborations grows faster than the number of authors, benefiting from the existing structure of the network. The temporal evolution also shows differences between well-established fields, such as Databases and Computer Architecture, and emerging fields, such as Bioinformatics and Geoinformatics. We also observed an increase on the average number of authors per paper for the three networks.