

Proceedings of the First International Workshop on
Internet Advertising: Theory and Practice
(IATP 2013)

August 1, 2013, Dublin, Ireland

held in conjunction with

SIGIR 2013

Workshop Organizers

Bin Gao
Jun Yan
Dou Shen
Tie-Yan Liu

Table of Contents

Workshop Description

Organizing Committee

Invited Talk 1: Financial Methods in Computational Advertising

Jun Wang

Invited Talk 2: Information Science versus Data Science for Digital Advertising and Marketing

James G. Shanahan

Invited Talk 3: Beyond Bag-of-words: Machine Learning for Matching

Jun Xu

Invited Talk 4: Semantically Related Bid Phrase Recommendation from Advertiser Web Pages

Sayan Pathak

Invited Talk 5: Large Scale Search Algorithms for Advertising Application

Vanja Josifovski

Invited Talk 6: Query-Ads Matching in Sponsored Search: Challenges and Solutions

Yunhua Hu

Invited Paper: Adaptive Keywords Extraction with Contextual Bandits for Advertising on Parked Domains

Shuai Yuan, Jun Wang, and Maurice van der Meer

Workshop Description

Internet advertising, a form of advertising that utilizes the Internet to deliver marketing messages and attract customers, has seen exponential growth since its inception around twenty years ago; it has been pivotal to the success of the World Wide Web. The dramatic growth of internet advertising poses great challenges to information retrieval, machine learning, data mining and game theory, and it calls for novel technologies to be developed.

The main purpose of this workshop is to bring together researchers and practitioners in the area of Internet Advertising and enable them to share their latest research results, to express their opinions, and to discuss future directions.

Internet advertising is one of the most important monetization engines of the Internet. It is highly associated with many IR applications such as web portals, search engines, social networks, e-commerce, mobile networks and apps, etc. Researchers in information retrieval, machine learning, data mining, and game theory are developing creative ideas to advance the technologies in this area.

We look forward to your contribution and attendance! See you in Dublin!

Organizing Committee

Bin Gao, Microsoft Research Asia

Jun Yan, Microsoft Research Asia

Dou Shen, Baidu Inc.

Tie-Yan Liu, Microsoft Research Asia

Invited Talk 1:

Financial Methods in Computational Advertising

Jun Wang

University College London, U.K.

Abstract: Computational Advertising has recently emerged as a new scientific sub-discipline, bridging the gap among the areas such as information retrieval, data mining, machine learning, economics, and game theory. In this tutorial, I shall present a number of challenging issues by analogy with financial markets. The key vision is that display opportunities are regarded as raw material “commodities” similar to petroleum and natural gas - for a particular ad campaign, the effectiveness (quality) of a display opportunity shouldn't rely on where it is brought and whom it belongs, but it should depend on how good it will benefit the campaign (e.g., the underlying web users' satisfactions or respond rates). With this vision in mind, I will go through the recently emerged real-time advertising, aka Real-Time Bidding (RTB), and provide the first empirical study of RTB on an operational ad exchange. We show that RTB, though suffering its own issue, has the potential of facilitating a unified and interconnected ad marketplace, making it one step closer to the properties in financial markets. At the latter part of this talk, I will talk about Programmatic Premium, i.e., a counterpart to RTB to make display opportunities in future time accessible. For that, I will present a new type of ad contracts, ad options, which have the right, but no obligation to purchase ads. With the option contracts, advertisers have increased certainty about their campaign costs, while publishers could raise the advertisers' loyalty. I show that our proposed pricing model for the ad option is closely related to a special exotic option in finance that contains multiple underlying assets (multi-keywords) and is also multi-exercisable (multi-clicks). Experimental results on real advertising data verify our pricing model and demonstrate that advertising options can benefit both advertisers and search engines.

Bio: Jun Wang is Senior Lecturer (Associate Professor) in University College London and Founding Director of MSc/MRes Web Science and Big Data Analytics. His main research interests are in the areas of information retrieval, data mining and online advertising. He has recently studied financial methods for online advertising. Dr. Wang has published over 70 research papers in leading journals and conference proceedings. He was a recipient of the Beyond Search – Semantic Computing and Internet Economics award sponsored by Microsoft Research, USA in 2007; he also received the Best Doctoral Consortium award in ACM SIGIR06 for his unified theory of collaborative filtering, the Best Paper Prize in ECIR09 for his pioneer work on Portfolio Theory of Information Retrieval, and the Best Paper Prize in ECIR12 for top-k retrieval modelling. Dr. Wang obtained his PhD degree in Delft University of Technology, the Netherlands; MSc degree in National University of Singapore, Singapore; and Bachelor degree in Southeast University, Nanjing, China.

Invited Talk 2:

Information Science versus Data Science for Digital Advertising and Marketing

James G. Shanahan

Independent Consultant, U.S.A.

Bio: Jimi has spent the last 25 years developing and researching cutting-edge information management systems that harness machine learning, information retrieval, and linguistics. During the summer of 2007, he started a boutique consultancy (Church and Duncan Group Inc., in San Francisco) whose major goal is to help companies leverage their vast repositories of data using statistics, machine learning, optimisation theory and data mining for big data applications (billions of examples) in areas such as web search, local and mobile search, and digital advertising and marketing. Church and Duncan Group's clients include Adobe, AT&T, Akamai, W3i, Digg.com, eBay, SearchMe.com, Ancestry.com, TapJoy, and SkyGrid.com. Along with architecting and developing large scale distributed statistical optimization systems for his clients, Jimi also leads and hires engineers and scientists, and provides business insights and strategic guidance to sales, analytics and business development groups within these organizations. In addition, Jimi has been affiliated with the University of California at Santa Cruz since 2009 where he teaches a sequence of graduate courses on big data analytics, machine learning, and stochastic optimization (TIM 206, ISM 209, ISM 250 and ISM251). He advises several high-tech startups (e.g., Quixey.com, NativeX, InferSystems) and is executive VP of science and technology at Irish Innovation Center (IIC). He has served as a fact and expert witness.

Prior to founding Church and Duncan Group Inc., Jimi was Chief Scientist and executive team member at Turn Inc. (an online ad network that has recently morphed to a demand side platform). Prior to joining Turn, Jimi was Principal Research Scientist at Clairvoyance Corporation where he led the "Knowledge Discovery from Text" Group. In the late 1990s he was a Research Scientist at Xerox Research Center Europe (XRCE) where he co-founded Document Souls, an anticipatory information system, where documents were given personalities of information services that foraged the web to stay informed and informative. In the early 90s, he worked on the AI Team within the Mitsubishi Group in Tokyo.

He has published six books, over 50 research publications, and 15 patents in the areas of machine learning and information processing. Jimi chaired CIKM 2008 (Napa Valley), co-chaired International Conference in Weblog and Social Media (ICWSM) 2011 in Barcelona, and was PC co-chair of ICWSM 2012 (Dublin). He co-chaired the ISSDM Workshop on Knowledge Management: Analytics and Big Data at UC Santa Cruz. He has organized several workshops in digital advertising as part of SIGIR, NIPS and SIGKDD. He is regularly invited to give talks at international conferences and universities around the world. Jimi received his Ph.D. in engineering mathematics from the University of Bristol, U. K. and holds a Bachelor of Science degree from the University of Limerick, Ireland. He is a Marie Curie fellow and member of IEEE and ACM. In 2011 he was selected as a member of the Silicon Valley 50 (Top 50 Irish Americans in Technology).

Invited Talk 3:

Beyond Bag-of-words: Machine Learning for Matching

Jun Xu

Huawei Noah's Ark Lab, Hong Kong

Abstract: Dealing with mismatch between query and ads (documents) is one of the most critical research problems in search. Recently researchers have spent significant effort to address the grand challenge. The major approach is to conduct more query and document understanding, and perform matching between enriched representations. With the availability of large amount of log data and advanced machine learning techniques, this becomes more feasible and significant progress has been made. In this talk, I will give a survey on newly developed machine learning technologies for matching. I will focus on the descriptions on the fundamental problems, as well as the novel solutions.

Bio: Jun Xu is Researcher at Noah's Ark Lab, Huawei Technologies in Hong Kong. He received his PhD in computer science from Nankai University China in 2006. He worked at Microsoft Research Asia during 2006 and 2012. He joined Huawei Noah's Ark Lab in 2012. Jun's research interest focuses on information retrieval and web search. He has published extensively in prestigious conferences and journals including SIGIR, WWW, WSDM, JMLR, and TOIS etc. Jun is very active in the research communities and served or is serving the top conferences and journals. He developed the learning to rank algorithms of IR-SVM and AdaRank, large scale topic models of RLSI and GMF. He released the source code of AdaRank, RLSI, and LETOR dataset to the academic.

Invited Talk 4:

Semantically Related Bid Phrase Recommendation from Advertiser Web Pages

Sayan Pathak

Microsoft, India

Abstract: Sponsored search systems such as Bing and Google has three players: advertisers, users and publishers. Users search for relevant information through search queries and advertisers bid for these user queries (bid keywords) to sell their services. As advertisers are not fully aware of the user queries, advertisers require keyword recommendation. And an important aspect of these keyword recommendation engines is to recommend relevant keywords for the given advertiser based on the advertiser's context such as website, business category etc. Previous work in the area of keyword recommendation is based on mining advertisers website and other context; however, these approaches are limited by the keywords present on the advertisers website. We propose the problem of keyword recommendation as a large-scale multi-label learning task where labels are keywords. In this talk, we present a multi-label random forest formulation that associates each data point (advertiser) with relevant subset of labels (keywords) from the universe of labels (keywords). Proposed approach automatically generates training data for the classifier from click logs without any human annotation or intervention. As the training complexity is polynomial in the number of training points and prediction complexity is logarithmic in the number of training points, proposed approach can be used for large-scale applications. Large-scale experiments conducted on MapReduce with 50 million webpages and 10 million keywords extracted from Bing logs reveal significant gains in P@10 compared to previous ranking and NLP based techniques.

Bio: Sayan Pathak received BS from the Indian Institute of Technology, Kharagpur, India in 1994, MS and PhD degrees in Bioengineering from the University of Washington, Seattle, in 1996 and 2000, respectively. Currently, he is working at the Online Services Division of Microsoft as a principal program manager leading algorithm R&D in the Bing Ads team. Prior to moving into the online advertising space, he has been researching and integrating cutting edge machine learning technology into Microsoft products in collaboration with Microsoft Research Labs both as a developer and program manager. He has been a faculty at the University of Washington for past 10 years and has over 8 years of teaching experience. Prior to joining Microsoft, he worked at Allen Institute for Brain Science, University of Washington and has been a principal investigator on several US National Institutes of Health (NIH) grants. He has published in leading journals such as Nature, Nature Neuroscience, IEEE and PNAS in the areas of large scale machine learning, computer vision, classification, image and signal processing.

Invited Talk 5:

Large Scale Search Algorithms for Advertising Application

Vanja Josifovski

Google, U.S.A.

Abstract: Today's online advertising comes in many different formats with different technology used for ad selection: from simple hash lookups to recommender systems. In this talk we give an overview of the different ad selection problems and show how they can be mapped to large scale, high dimensional similarity search. Specifically we will overview the WAND algorithm which was proposed about a decade ago for high-dimensional similarity search and has thereafter been adapted for multiple search and data mining applications related to online advertising.

Bio: Vanja Josifovski is a Technical Lead in the Strategic Technologies group at Google where he is leading projects in the areas of recommender systems, information extraction and computational advertising. Prior to Google, Vanja was a Sr. Director of Research at Yahoo! Research where he worked on the next generation Contextual Advertising, Sponsored Search and Display Advertising Targeting platforms. Based on his work, Vanja co-taught a computational advertising course at Stanford University. Even earlier in his career, Vanja was a Research Staff Member at the IBM Almaden Research Center working on databases and enterprise search.

Invited Talk 6:

Query-Ads Matching in Sponsored Search: Challenges and Solutions

Yunhua Hu

Alibaba Corporation, China

Abstract: In sponsored search, usually all related Ads of a given query are matched by following a three-stage matching model. The model consists of query rewriting stage, Bid-Ads ranking stage, and Ads ranking stage. However, such a three-stage matching model is still not perfect for sponsored search. For instances, it is not easy to evaluate the accuracy of the whole query-ads matching process. It is also difficult to optimize each stage. It is even not clear whether there is still space to be improved. In this session, I will introduce the challenges we faced. I will also share some initial solutions for these challenges.

Bio: Yunhua Hu is a Senior Technical Specialist from Alimama business group in Alibaba Corporation. He focuses on query analysis based on big data, personalized search, matching and ranking algorithm in search advertising, etc. Before he joined in Alibaba, he had been worked in Natural Language Processing (NLP) group and Web Search and Mining(WSM) group in Microsoft Research Asia (MSRA) for more than 5 years. He conducted some advanced research on Enterprise Search, Academic Search, and log mining for Web Search. Some key technologies have been transferred into Microsoft Office and Bing search engine. He also published several papers in SIGIR, WSDM, CIKM, JCDL, and AAAI, etc.

Invited Paper:

Adaptive Keywords Extraction with Contextual Bandits for Advertising on Parked Domains

Shuai Yuan

University College London, U.K.

Jun Wang

University College London, U.K.

Maurice van der Meer

B.V. DOT TK, Netherlands

Abstract: Domain name registrars and URL shortener service providers place advertisements on the parked domains (Internet domain names which are not in service) in order to generate profits. As the web contents have been removed, it is critical to make sure the displayed ads are directly related to the intents of the visitors who have been directed to the parked domains. Because of the missing contents in these domains, it is non-trivial to generate the keywords to describe the previous contents and therefore the users intents. In this paper we discuss the adaptive keywords extraction problem and introduce an algorithm based on the BM25F term weighting and linear multi-armed bandits. We built a prototype over a production domain registration system and evaluated it using crowdsourcing in multiple iterations. The prototype is compared with other popular methods and is shown to be more effective.

Adaptive Keywords Extraction with Contextual Bandits for Advertising on Parked Domains

Shuai Yuan, Jun Wang
Department of Computer Science,
University College London
London, United Kingdom
{s.yuan, j.wang}@cs.ucl.ac.uk

Maurice van der Meer
B.V. DOT TK
Amsterdam, Netherlands
maurice@dot.tk

ABSTRACT

Domain name registrars and URL shortener service providers place advertisements on the parked domains (Internet domain names which are not in service) in order to generate profits. As the web contents have been removed, it is critical to make sure the displayed ads are directly related to the intents of the visitors who have been directed to the parked domains. Because of the missing contents in these domains, it is non-trivial to generate the keywords to describe the previous contents and therefore the users intents. In this paper we discuss the adaptive keywords extraction problem and introduce an algorithm based on the BM25F term weighting and linear multi-armed bandits. We built a prototype over a production domain registration system and evaluated it using crowdsourcing in multiple iterations. The prototype is compared with other popular methods and is shown to be more effective.

1. INTRODUCTION

Online advertising is a fastest growing area in IT industry in recent years. In an online advertising eco-system, yield optimisation is the core problem of the supply side, aka, publishers, and is commonly dealt with by managing revenue channels [21]; selecting relevant ads [28]; optimising reserve prices [19]; controlling advertising level [10], etc. The work presented in this paper is focused on a specific problem: advertising on parked domains. Quite often, domains hold valid web contents until certain stages when they are bought/revoked/suspended/expired later. Before those domains are (re)used, they are empty with no content displayed but ads. This occurs commonly in the URL shortener services and domain registration services. As people may not know the webpages do not exist any more, there are still many in-links available online which generate good amount of traffic towards the parked domains. As a general practice, the domain providers or registrars host display ads on these parked domains in order to generate profit, c.f. Figure 1. Ideally the ads are required to be as relevant as possible to the visitors. A clear difference from the existing contextual advertising research [27, 6] is that the context (webpage content) is no longer available when impressions are created and therefore it is hard to generate keywords to describe potential visitors (interests or intents).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IATP 2013, August 1, 2013, Dublin, Ireland.
Copyright is held by the author/owner(s).

In this paper, we use contextual multi-armed bandits [16, 17] to predict the relevant keywords over time. Different from the standard multi-armed bandits, every arm is associated with a feature vector as the side information known to the player. We employ the general linear model (GLM) to describe the relationship of rewards (relevance) and features (keywords), where the model may also be referred to as the linear bandits [11]. We take the upper confidence bound (UCB) [4] approach to explore and exploit optimal arms. Moreover, we applied crowdsourcing to collect user feedback and adjust the algorithm. We chose accessors randomly from `oDesk.com`, one of the largest online crowdsourcing services, given they had the sufficient knowledge to understand the webpages and their tasks. Besides screenshots and remote desktop monitoring, we embedded multiple traps in judge items and considered the user failed the task if 30% traps were triggered. To validate the results, we employed both Fleiss' Kappa [14] and Cohen's Kappa [8] to measure the agreement of judgements. Our experiments show the effectiveness of the proposed approach in addressing the problem.

2. RELATED WORKS

There are commercial tools generating high co-occurrence or similar phrases as suggestions based on seed terms provided by user, like Google AdWords Keyword Tool¹. There are two problems for these tools: 1) it still requires fair amount of manual work to input seeds and choose from suggestions. 2) the topics of generated phrases are based on user query logs, existing bid phrases, and lexical analysis, and may easily drift from the original from webpages.

Keywords extraction is largely considered a supervised learning problem [24, 13, 26, 20] where the algorithms learn to classify as positive or negative examples of keywords based on training sets. These algorithms need expensive human labelled dataset in advance, and usually perform the learning process offline. The model could become inaccurate when users' interests change over time. In [13] linguistic knowledge is introduced such as noun-phrase-chunks (NP-chunks) and part-of-speech (POS) to outperform using statistics features only. These linguistic features are also used in our work. Besides, query logs are also used as a good reflect of users' interests [12].

There are also extensive research about keywords suggestion. In [7] the keywords are suggested based on concept hierarchy mapping therefore the suggestions are not limited to the bag-of-words of the webpage, and may expand to non-obvious ones which are categorized in bigger concepts. The work of [15, 3] recognise that the bid prices for hot keywords are high therefore would cost more, and try to find related

¹adwords.google.com/select/KeywordToolExternal

non-obvious keywords that are cheaper. Although these keywords may have lower traffic, but when combined the traffic could match that of a hot one, while these keywords cost still less.

The work of [27] proposed a classifier that uses multiple text features, including how often the term occurs in search query logs, to extract keywords for ads targeting based on logistic regression. The system discussed in [20] first generates candidates by several methods including a translation model capable of generating phrases not appearing in the text of the pages. Then candidates are ranked in a probabilistic framework using both the translation model favouring relevant phrases, as well as a language model favouring well-formed phrases. Another relevant work can be found in [23]. The authors proposed an exploration-exploitation algorithm of sorting keywords in an descending order of profit-to-cost ratio and adaptively identify the set of keywords to bid on based on historical performance, with a daily budget constraint. In this paper we try to find the keywords from the given webpage with additional web knowledge, rather than find profitable ones from a very large set (like 50k). Then we leave the matching between keywords and ads to display networks/exchanges, such as Google AdSense.

In [18] the authors propose a combination algorithm of using upper confidence bound (UCB) and ϵ -greedy to solve the exploration-exploitation dilemma. Their target is to select high profit ads which would be fed in contextual advertising platforms. The feature vectors of ads are not used in their system as side information, instead they use standard bandits considering the reward following an unknown stochastic process. Our research is partially inspired by their work.

3. ONLINE KEYWORDS EXTRACTION

From the publishers' perspective the process of extracting keywords is illustrated in Figure-2. It is an iterative process of selecting candidates and update the model according to the feedback. This task of extracting keywords against user feedback is directly linked to the contextual multi-armed bandits problem. The linear multi-armed bandits used in our work [9, 22, 2, 4] is a special case of contextual bandits. It has the same setting with standard bandits except the availability of the side information. The side information of each arm determines the reward for pulling the arm therefore could be used to make decisions. In our work, we mainly exploited text features from webpages under domains being evaluated, including TF-IDF scores in title, content, keyword and description in HTML meta tag, header, anchor text of in-links, and part of speech. Since we hold the webpages or redirections of parked domains, extracting these features is possible.

First of all we define reward of iteration step $t \in [0, T]$ as a real number in $[0, 1]$,

$$r(t) \in [0, 1] \quad (1)$$

The reward could be in various form, for example 1) relevance scores of selected keywords against the webpage; 2) the clickthrough rates (CTR) of ads, or 3) profit gained in each iteration. Apparently the relevance scores, CTRs and profit are linked, however not guaranteed with any kind of linear/non-linear relationship to our best extent of knowledge. In our work we used human judged relevance score for simplicity, also due to the difficulty of acquiring necessary data of CTR and profit. Besides, we employed the crowd-sourcing approach to get the feedback in a cheap and fast way. We considered each webpage a K-armed bandit machine and every arm a word (or a multi-word phrase). In

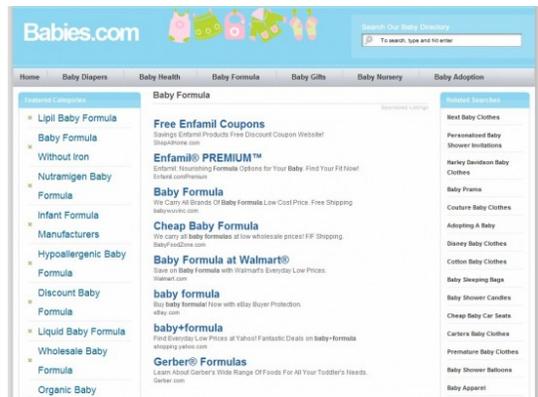


Figure 1: An example of advertising on parked domains. Instead of showing an error or *under construction*, relevant ads are displayed.

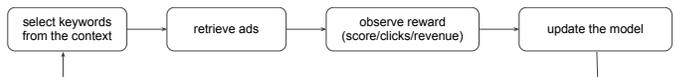


Figure 2: The keywords extraction task could be an iterative process. The publisher would like to exploit current optimal keywords as well as to explore potentially better ones.

each iteration an arm will be pulled resulting in selection of the corresponding word/phrase to be the keywords of the webpage. Different from standard bandits, each arm ($i \in K$) of the linear bandits is associated with some D -dimension feature vector $\mathbf{x}_i \in \mathbb{R}^{D \times 1}$ which is already known. The expected reward (user feedback) is given by the inner product of its feature vector \mathbf{x}_i and some fixed, but initially unknown parameter (column) vector \mathbf{w} . That is, the reward is a linear function of the feature vector and unknown parameters,

$$r_i = \mathbf{x}_i' \cdot \mathbf{w} \quad (2)$$

The goal here is to get the optimal reward. We take the LINREL algorithm [4] highlighting upper confidence bound (UCB) to deal with the exploration-exploitation dilemma.

3.1 Upper Confidence Bound Approach

The idea of LINREL algorithm is to estimate the reward for the i -th arm from the linear combination of historical reward received, with a high probability. We expand the LINREL algorithm a bit to help the paper self-contained. First we write the feature vector of the i -th arm as the linear combination of the previous chosen feature vectors (this is always possible except for the initial d iteration steps),

$$\mathbf{x}_i(t) = \mathbf{X}(t) \cdot \mathbf{a}_i(t) \quad (3)$$

where $\mathbf{a}_i(t) \in \mathbb{R}^{t \times 1}$ is the coefficient of the linear combination and $\mathbf{X}(t)$ is the feature matrix selected in past iteration steps with dimension $D \times t$,

$$\mathbf{X}(t) = [\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t-1)] \quad (4)$$

where the $\mathbf{x}(t)$ denotes the feature vector used in t iteration step (without specifying the selected arm). With the same coefficient the reward could be written as,

$$r_i(t) = \mathbf{x}_i(t)' \cdot \mathbf{w} = (\mathbf{X}(t) \cdot \mathbf{a}_i(t))' \cdot \mathbf{w} = \mathbf{R}(t)' \cdot \mathbf{a}_i(t) \quad (5)$$

where $\mathbf{R}(t) \in \mathbb{R}^{t \times 1}$ is a vector of historical reward,

$$\mathbf{R}(t) = [r(1), r(2), \dots, r(t-1)]' \quad (6)$$

This gives a good estimate $\mathbf{R}(t) \cdot \mathbf{a}_i(t)$ for $r_i(t)$. The algorithm keeps the variance small to maintain a narrow confidence interval of the estimate. By assuming i.i.d of $r_i(t)$ (which is true for our keywords extraction task) and since $r_i(t) \in [0, 1]$ the variance of this estimate is bounded by $\|\mathbf{a}_i(t)\|^2/4$. In order to get $\mathbf{a}_i(t)$ first calculate the eigenvalue decomposition,

$$\mathbf{X}(t) \cdot \mathbf{X}(t)' = \mathbf{U}(t)' \cdot \mathbf{\Delta}[\lambda_1, \lambda_2, \dots, \lambda_d] \cdot \mathbf{U}(t) \quad (7)$$

where $\lambda_1, \dots, \lambda_k \geq 1$ and $\lambda_{k+1}, \dots, \lambda_d < 1$. Then for each feature vector $\mathbf{x}_i(t)$ write,

$$\mathbf{z}_i(t) = \mathbf{U}(t) \cdot \mathbf{x}_i(t) = [x_{i,1}(t), x_{i,2}(t), \dots, x_{i,d}(t)]' \quad (8)$$

$$\mathbf{u}_i(t) = [x_{i,1}(t), \dots, x_{i,k}(t), 0, \dots]' \quad (9)$$

$$\mathbf{v}_i(t) = [0, \dots, x_{i,k+1}(t), \dots, x_{i,d}(t)]' \quad (10)$$

The coefficient $\mathbf{a}_i(t)$ is calculated as,

$$\mathbf{a}_i(t) = \mathbf{u}_i(t)' \cdot \mathbf{\Delta}[\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_k}, \dots, 0] \cdot \mathbf{U}(t) \cdot \mathbf{X}(t) \quad (11)$$

Then with probability of $1 - \delta/T$ the expected reward of i -th arm is,

$$ucb_i(t) = \mathbf{R}(t) \cdot \mathbf{a}_i(t) + \sigma \quad (12)$$

where σ is the width of the confidence bounds by using Azuma-Hoeffding bound [5],

$$\sigma = \|\mathbf{a}_i(t)\|(\sqrt{\ln(2TK/\delta)} + \|v_i(t)\|) \quad (13)$$

where δ is used to confine a narrow confidence interval and recall K is the number of arms. The arm with highest $ucb_i(t)$ score will be selected. Apparently the arm got selected because the combination of its expected reward ($\mathbf{R}(t) \cdot \mathbf{a}_i(t)$) or potential reward (σ) is high. The former leads to exploitation and the latter results in exploration.

4. EXPERIMENTS AND RESULTS

In this section we introduce the architecture of the prototype system and its evaluation against BM25F [29], KEA [25], and two popular web services.

4.1 Prototype System

We have built a prototype system RAZORCLAW according to the model discussed above. The system is open source, written in Java, and online running. Its architecture is illustrated in Figure-3. Generally the prototype system is made of 3 main parts: the crawler, parser, and the ranker. In the parsing engine, different modules will be used for different languages.

In the crawling module, the RAZORCLAW first loads meta information from the parked domain database, including the previously displayed webpage, the anchor texts and referrer URLs collected from the Internet. The webpage is crawled, too. The next step is to detect the correct encoding of the text (sometimes not presented in HTML) and the language. The language detector we use² is based on naive Bayesian filter and has reported 99% precision over 49 languages.

Once the encoding and language are detected we convert the content to UTF-8 and invoke corresponding NLP processor. We employ openNLP³ for major western languages and IKAnalyzer⁴ for Chinese-Japanese-Korean-Vietnam languages. The language support is not the main point of the paper

²code.google.com/p/language-detection

³incubator.apache.org/opennlp

⁴code.google.com/p/ik-analyzer

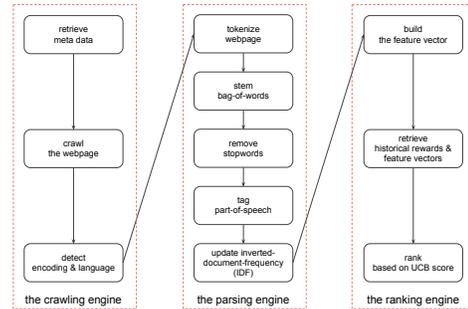


Figure 3: The architecture of the RAZORCLAW prototype system. We divide the system into 3 main parts. In every part there are loosely bounded modules with each for a single task. These modules could be changed or updated flexibly. We intent to create an open framework for the keywords extraction task.

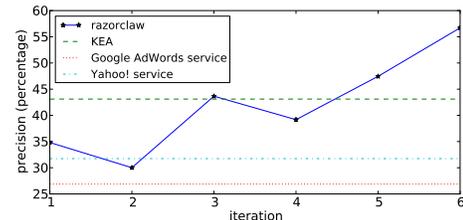


Figure 4: The performance comparison of competing algorithms and services.

however our system processes more than 50 languages which is especially useful for free domain services, where we discovered a great portion of registrations are from south east Asia, India, and Arabic countries.

After stemming, removing stopwords, part-of-speech (POS) tagging and saving the inverted-document-frequency (IDF), each phrase in the bag-of-words is processed to generate its local feature vector and to retrieve historical features and rewards. According to the ranking result our system gives candidate keywords (usually the top 3).

4.2 Datasets and Algorithms

We used 165 domains from Dot.tk's database and collected all possible side information. We compared our prototype system with BM25F algorithm [29], KEA algorithm⁵ [25], Yahoo! Term Extraction service⁶ and Google Targeting Idea service⁷.

The parameters of BM25F algorithm was obtained from [29]. KEA is a famous keywords extraction tool in academic research. It is capable of extraction keywords with domain specific knowledge, for example with Wikipedia article names. However in our experiments we select domains randomly and do not restrict the result to any specific area.

Both web services require registration as a developer and incur cost if exceeding the free quota. According to the specification, we send text content of webpages (including the title and meta) to the Yahoo! service and send URLs to the Google one.

4.3 Crowdsourcing Based Experiments

In order to carry out the evaluation in an efficient and cheap fashion, we employed a crowdsourcing platform to judge the results of multiple algorithms. We conducted a test on the popular crowdsourcing platform oDesk.com and ran-

⁵www.nzdl.org/kea

⁶goo.gl/KXaPw

⁷goo.gl/CpNgp

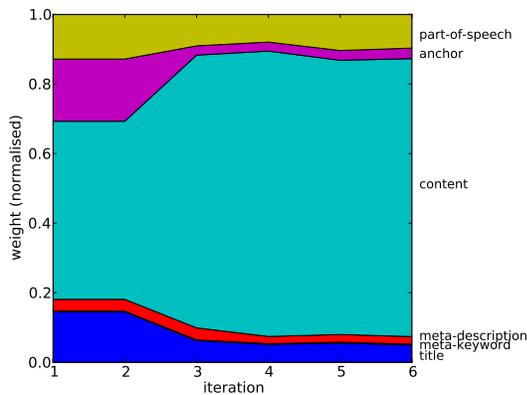


Figure 5: The evolution of weights of some features of razorclaw system. Features with too small weights are not included here. The initial weights were obtained from [29]. The result showed that the *content* of webpage is the most influential factor, followed by *part-of-speech* and *title*. The *keyword* and *description* in the HTML meta tag are in fact not trustworthy.

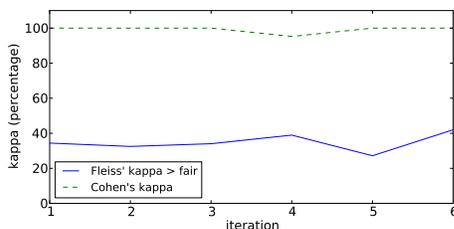


Figure 6: The plot of agreement measurement for crowdsourcing ranking results. The *fair* in Fleiss' kappa corresponds to 0.2

domly selected 23 human assessors who passed the test. We then asked them to rank the relevance between the domain and keywords on a 1 (least relevance) to 5 (most relevance) scale. We gave very detailed instructions to help keep the ranking as consistent as possible among different assessors and in different iteration steps. For broken links or empty keywords (some algorithm failed to give the result) the assessors were asked to give a 0. We chose oDesk.com for its simplicity, openness, big candidate pool, and various tools to monitor the progress. We designed traps, used screenshots and remote desktop monitoring to reduce fraud and malicious behaviours. Among the returns 3 users were marked careless (by triggering 3 out of 10 traps) and their results were completely removed from the pool.

4.4 Results

We carried out 6 iterations of experiments in total, on 7 June, 17 July, 4 August, 11 August, 15 August, and 20 August 2011. The comparison of competing algorithms is reported in Figure 4. As introduced above, the relevance ranking ranges from 0-5 with 0 is specifically for failures (website did not open, encoding or language did not recognised, or other issues resulting in failure of extracting keywords). In every iteration each domain was ranked by at least 2 assessors.

First we compare the overall precision (cumulative relevance ranking against the highest possible score). From Figure 4 we can see the RAZORCLAW performed significantly better than the 2nd best (KEA), with 8.29% behind in the first iteration, but 13.6% improvement in the last one. The web services performed less satisfying mainly due to they were limited to English language at the time of test.

From the precision plot the improvement of RAZORCLAW is clearly observable. At iteration 2 and 4 the algorithm

performed worse than the previous round, suffering from the exploration. However, the algorithm was able to pick up new sets of weights of features as anticipated. From Figure 5 we can see the corresponding evolution of the weights. In the figure we listed top 6 features and did not include ones with too small weights. The *content* was the most influential factor, followed by *part-of-speech* and *title*. The *keyword* from the HTML meta tag is in fact not trustworthy, which is consistent with the no-use decision in major search engines like Google [1]. The measurement of agreement is plotted in Figure 6. We computed the Cohen's kappa of any two assessors and reported its average of each iteration. Similarly the Fleiss' kappa of all assessors of each iteration was plotted. Generally speaking the agreement of results was satisfying.

5. CONCLUSION

In this paper, we discussed an adaptive keyword extraction algorithm based on BM25F and contextual multi-armed bandits. We showed its good performance in crowdsourcing based experiments, and reported interesting findings including the evolution of weights. We plan to include more features and conduct larger scale of evaluation in future.

6. REFERENCES

- [1] Meta tags. goo.gl/iTgs9 (last visited 20/6/2013).
- [2] Y. Abbasi-Yadkori, A. Antos, and C. Szepesvári. Forced-exploration based algorithms for playing in stochastic linear bandits. In *Proceedings of the COLT 2009*.
- [3] V. Abhishek and K. Hosanagar. Keyword generation for search engine advertising using semantic similarity between terms. In *Proceedings of the ACM EC 2007*.
- [4] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 2003.
- [5] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 1967.
- [6] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *Proceedings of the ACM SIGIR 2007*.
- [7] Y. Chen, G.-R. Xue, and Y. Yu. Advertising keyword suggestion based on concept hierarchy. In *Proceedings of the ACM WSDM 2008*.
- [8] J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 1960.
- [9] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the COLT 2008*.
- [10] R. M. Dewan, M. L. Freimer, and J. Zhang. Management and valuation of advertisement-supported web sites. *Journal of Management Information Systems*, 2003.
- [11] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 2010.
- [12] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal. Using the wisdom of the crowds for keyword generation. In *Proceeding of the ACM WWW 2008*.
- [13] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the EMNLP 2003*.
- [14] F. L. Joseph. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 1971.
- [15] A. Joshi and R. Motwani. Keyword generation for search engine advertising. In *Proceedings of the ICDM 2006*.
- [16] J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in Neural Information Processing Systems*, 2007.
- [17] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the ACM WWW 2010*.
- [18] W. Li, X. Wang, R. Zhang, Y. Cui, J. Mao, and R. Jin. Exploitation and exploration in a performance based contextual advertising system. In *Proceedings of the ACM SIGKDD 2010*.
- [19] M. Ostrovsky and M. Schwarz. Reserve prices in internet advertising auctions: A field experiment. 2009.
- [20] S. Ravi, A. Broder, E. Gabrilovich, V. Josifovski, S. Pandey, and B. Pang. Automatic generation of bid phrases for online advertising. In *Proceedings of the ACM WSDM 2010*.
- [21] G. Roels and K. Fridgeirsdottir. Dynamic revenue management for online display advertising. *Journal of Revenue & Pricing Management*, 2009.
- [22] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 2010.
- [23] P. Rusmevichientong and D. P. Williamson. An adaptive algorithm for selecting profitable keywords for search-based advertising services. In *Proceedings of the ACM EC 2006*.
- [24] P. D. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2000.
- [25] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning. Kea: Practical automatic keyphrase extraction. In *Proceedings of the ACM DL 1999*.
- [26] H. Wu, G. Qiu, X. He, Y. Shi, M. Qu, J. Shen, J. Bu, and C. Chen. Advertising keyword generation using active learning. In *Proceedings of the ACM WWW 2009*.
- [27] W.-t. Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the ACM WWW 2006*.
- [28] S. Yuan and J. Wang. Sequential selection of correlated ads by pomdps. In *Proceedings of the ACM CIKM 2012*.
- [29] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft cambridge at trec-13: Web and hard tracks. In *Proceedings of the TREC 2004*. Citeseer, 2004.