Proceedings

Proceedings of the 14th Dutch-Belgian Information Retrieval Workshop

Preface

This volume contains the papers presented at DIR 2015: 14th Dutch-Belgian Information Retrieval Workshop held on November 27, 2015 in Amsterdam.

The committee decided to accept 12 submissions for oral and poster presentation, 14 for poster presentation, 8 for demo presentation and 4 as lightning talks. Further, the committee decided to invite two keynote speakers. Each submission was reviewed by 3 program committee members.

The primary aim of the Dutch-Belgian Information Retrieval (DIR) workshop is to provide an international meeting place where researchers from the domain of information retrieval and related disciplines can exchange information and present innovative research developments.

November 21, 2015
AMSTERDAM

Artem Grotov
Christophe Van Gysel
Evangelos Kanoulas
Hosein Azarbonyad
Nikos Voskarides
Petra Best
Xinyi Li

Table of Contents

Keynotes

Designing Human Feedback Data for Machine Learning 1
Thorsten Joachims

Should we optimize search engines for social and personal welfare? 2
Elad Yom-Tov

Abstracts

Determining the Presence of Political Parties in Social Circles 3
Christophe Van Gysel, Bart Goethals and Maarten de Rijke

Short Text Similarity with Word Embeddings 4
Tom Kenter and Maarten de Rijke

Early Detection of Topical Expertise in Community Question Answering . 5
David van Dijk, Manos Tsagkias and Maarten de Rijke

Summarizing Contrastive Themes via Hierarchical Non-Parametric Processes 6
Zhaochun Ren and Maarten de Rijke

Multileave Gradient Descent for Fast Online Learning to Rank 7
Anne Schuth, Harrie Oosterhuis, Shimon Whiteson and Maarten de Rijke

Predicting Relevance based on Assessor Disagreement: Analysis and Practical Applications for Search Evaluation 8
Thomas Demeester, Robin Aly, Djoerd Hiemstra, Dong Nguyen and Chris Develder

Time-Aware Authorship Attribution of Short Texts 9
Hosein Azarbonyad, Mostafa Dehghani, Maarten Marx and Jaap Kamps

Time-aware Personalized Query Auto Completion 10
Fei Cai and Maarten de Rijke

Learning to Explain Entity Relationships in Knowledge Graphs 11
Nikos Voskarides, Edgar Meij, Manos Tsagkias, Maarten de Rijke and Wouter Weerkamp

Lost but Not Forgotten: Finding Pages on the Unarchived Web 12
Hugo Huurdeman, Jaap Kamps, Thaer Samar, Arjen de Vries, Richard Rogers and Anat Ben David

Behavioral Dynamics from the SERP's Perspective: What are Failed SERPs and How to Fix Them? 13
Julia Kiseleva, Jaap Kamps, Vadim Nikulin and Nikita Makarov

Categorizing Events using Spatio-Temporal and User Features from Flickr 14
Steven Van Canneyt, Steven Schockaert and Bart Dhoedt

CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data 15
Anca Dumitrache, Oana Inel, Benjamin Timmermans, Lora Aroyo and Robert-Jan Sips

Dynamic Collective Entity Representations for Entity Ranking 16
David Graus, Manos Tsagkias, Wouter Weerkamp, Edgar Meij and Maarten de Rijke

Mining, Ranking and Recommending Entity Aspects 17
Ridho Reinanda, Edgar Meij and Maarten de Rijke

Eye-tracking Studies of Query Intent and Reformulation 18
Carsten Eickhoff, Sebastian Dungs and Tuan Vu Tran

From Multistage Information-Seeking Models to Multistage Search Systems 19
Hugo C. Huurdeman and Jaap Kamps

Struggling and Success in Web Search 20
Daan Odijk, Ryen White, Ahmed Hassan Awadallah and Susan Dumais

What Makes Book Search in Social Media Complex? 21
Marijn Koolen, Toine Bogers, Antal van Den Bosch and Jaap Kamps

Translation Model Adaptation Using Genre-Revealing Text Features 22
Marlies van der Wees, Arianna Bisazza and Christof Monz

Image2Emoji: Zero-shot Emoji Prediction for Visual Media 23
Spencer Cappallo, Thomas Mensink and Cees Snoek

Learning to Combine Sources of Evidence for Indexing Political Texts . . . 24
Mostafa Dehghani, Hosein Azarbonyad, Jaap Kamps and Maarten Marx

Automatically Assessing Wikipedia Article Quality by Exploiting Article-Editor Networks 25
Xinyi Li and Maarten de Rijke

A Hybrid Approach to Domain-Specific Entity Linking 26
Alex Olieman, Jaap Kamps, Maarten Marx and Arjan Nusselder

Using Logged User Interactions for Ranker Evaluation 27
Artem Grotov, Shimon Whiteson and Maarten de Rijke

KISS MIR: Keep It Semantic and Social Music Information Retrieval 28
Amna Dridi

Demonstrations

A Search Engine with Reading Level Specific Results 29
Thijs Westerveld

QUINN: Query Updates for News Monitoring 30
Suzan Verberne, Thymen Wabeke and Rianne Kaptein

BioMed Xplorer: A tool for exploring biomedical knowledge 31
Mohammad Shafahi, Hamideh Afsarmanesh and Hayo Bart

Let the Children Play - Developing a Child Feedback Collection System for Text Readability Assessments 32
Rutger Varkevisser, Theo Huibers and Thijs Westerveld

Multilingual Word Embeddings from Sentence Representations 33
Benno Kruit and Sara Veldhoen

Knowledge Discovery in Medical Forums 34
Erik Boertjes, Rianne Kaptein, Martijn Spitters and Wessel Kraaij

Implementation of Specialized Product Search Features In a Large-Scale Search And Merchandising Solution 35
Ivan Zamanov, Andreas Brückner and Raul Leal

Self-Learning Search Suite 36
Manos Tsagkias and Wouter Weerkamp

Lightning talks

Open Search 38
Anne Schuth

Retrieving Research Trends in Twitter 39
Amna Dridi

The Big Data Fad 40
Jeroen Bulters

Exact Match in IR 41
Arjen de Vries

Program Committee

Robin Aly	University of Twente
Marc Bron	Yahoo Labs
Davide Ceolin	VU University Amsterdam
Aleksandr Chuklin	University of Amsterdam
Victor de Boer	VU Amsterdam
Thomas Demeester	Ghent University
Chris Develder	Ghent University - iMinds
Carsten Eickhoff	ETH Zurich
Antske Fokkens	VU Amsterdam
David Graus	University of Amsterdam
Artem Grotov	University of Amsterdam
Claudia Hauff	Delft University of Technology
Jiyin He	CWI
Hugo Huurdeman	University of Amsterdam
Evangelos Kanoulas	University of Amsterdam
Rianne Kaptein	TNO
Tom Kenter	University of Amsterdam
Mike Kestemont	CLiPS
Julia Kiseleva	University of Saint-Petersburg
Marijn Koolen	University of Amsterdam
Xinyi Li	University of Amsterdam
Ilya Markov	University of Amsterdam
Danish Nadeem	University of Twente
Dong Nguyen	University of Twente
Daan Odijk	University of Amsterdam
Tobias Schnabel	Cornell University
Kim Schouten	Erasmus University Rotterdam
Anne Schuth	University of Amsterdam
Adith Swaminathan	Cornell University
Steven Van Canneyt	Ghent University - iMinds
Christophe Van Gysel	University of Amsterdam
Suzan Verberne	Institute for Computing and Information Sciences, Radboud University Nijmegen
Ivan Vulić	Department of Computer Science, KU Leuven
Jeroen Vuurens	Delft University of Technology

Designing Human Feedback Data for Machine Learning

Thorsten Joachims
Cornell University
tj@cs.cornell.edu

Abstract

Machine Learning has become one of the key enabling methods for building information access systems. This is evident in search engines, recommender systems, and electronic commerce, while other applications are likely to follow in the near future. In these systems, machine learning is not just happening behind the scenes, but it is increasingly used to directly interact with human users. In fact, much of the Big Data we collect today are the decisions that people make when they use these human-interactive learning systems we built.

In this talk, I argue that for building human-interactive learning system it is crucial to not only design the learning algorithm, but also to design the mechanism for generating the data from the human users. Towards this goal, the talk explores how integrating microeconomic models of human behavior into the learning process leads to new learning models that no longer misrepresent the user as a labeling subroutine. This motivates an interesting area for research in information retrieval and machine learning, with connections to rational choice theory, econometrics, and behavioral economics.

Should we optimize search engines for social and personal welfare?

Elad Yom-Tov
Microsoft Research
eladyt@microsoft.com

Abstract

Internet search engines are traditionally optimized to rank highest the most relevant results, that is, those results that satisfy the users information need. Information need and whether it was met is often judged by users themselves. However, this definition of relevance is subjective to users understanding of their information need. In my talk I will show that other notions of relevance are possible, notions that take social or personal welfare into account, and will ask whether we should adopt these alternative relevance measures.

In my talk I will discuss our results in promoting civil discourse through search engine diversity. I will also show that when searching for health information, results deemed relevant by traditional search engines can entrench people in understanding which can lead to unhealthy behaviors or to misunderstanding of ones medical condition. I will argue that in all these cases, a wider notion of relevance might be required.

I will end with a discussion of the pros and cons of adopting a wider definition of relevance.

Determining the Presence of Political Parties in Social Circles (Abstract)*

Christophe Van Gysel
University of Amsterdam
cvangysel@uva.nl

Bart Goethals
University of Antwerp
bart.goethals@uantwerp.be

Maarten de Rijke
University of Amsterdam
derijke@uva.nl

ABSTRACT

We derive the political climate of the social circles of Twitter users using a weakly-supervised approach. By applying random walks over a sample of Twitter's social graph we infer a distribution indicating the presence of eight Flemish political parties in users' social circles in the months before the 2014 elections. The graph structure is induced through a combination of connection and retweet features and combines information of over a million tweets and 14 million follower connections. We solely exploit the social graph structure and do not rely on tweet content. For validation we compare the affiliation of politically active Twitter users with the most-influential party in their network. On a validation set of 700 politically active individuals we achieve F_1 scores of 0.85 and greater.

INTRODUCTION

Blogs and social networks play an important role in the diffusion of political ideas [1]. We investigate the spread of influence of political parties in Twitter. We look at the contributions of eight political parties in the months before the 2014 elections in Flanders. We postulate that political influences travel through social graphs similarly to how ideas spread viva voce. Many social networks exhibit homophilic properties [4], implying that personal networks are grounded in sociodemographic, behavioral and intrapersonal characteristics [5]. Based on these properties we extract link-based features from interaction data on Twitter and simulate a random walk over an induced graph structure.

METHODOLOGY

Classification of nodes is performed through an iterative algorithm [2, 3] equivalent to performing random walks over an absorbing Markov chain. Their formulation is similar to the iterative formulation of the PageRank algorithm [6].

EXPERIMENTS

We consider a node classification problem over eight political parties active in the Flemish region of Belgium. While some of these parties share common views, they are relatively spread out over the political spectrum. These parties collectively published lists of 780 Twitter users with whom they associate themselves. We consider these lists as ground truth data and use these to measure classification performance and to determine the interaction weights.

We targeted 12 254 Twitter users who followed at least two of these eight parties and consequently retrieved their information as described in the previous section. We gathered 1 249 091 tweets (of which 273 213 were retweets) and 14 849 213 follower connections. These connections and tweets referred to at least 10 million users in total, which corresponds to the total number of nodes $|V|$.

We then built a graph structure from the gathered data. More precisely we introduce a directed edge from user i to user j if user i follows user j or when user i retweets a message from user j. In an initial experiment we assign both these interactions a unit weight. Later we also considered weights for the inversed edges, such that if user i follows user j a directed edge is added from user j to user i and similarly for retweets. We also considered a weight that is added when both user i and user j follow each other, which indicates reciprocal following. These additional interactions are interesting as they give insight in how actions of other users influences one's position in the social graph.

We asked Twitter users to evaluate their individual classification a week before the Flemish elections of 2014. Users were shown a distribution over the eight political parties and could voluntarily provide feedback between 0 and 100, where a higher score indicates a better classification. Some users expressed concern as they observed a non-zero probability for right- or left-winged extremist parties. We believe that some feedback scores were purposely negative so as to deny association with these parties, even though these associations were negligible in a statistical setting. We received feedback from 2 258 users. The distribution is characterized by its population mean (51.57), standard deviation (35.97) and median (62.0) [7].

REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *LinkKDD '05*, pages 36–43. ACM, 2005.

[2] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for Youtube: Taking random walks through the view graph. In *WWW 2008*, pages 895–904, 2008.

[3] S. Bhagat, G. Cormode, and S. Muthukrishnan. Node classification in social networks. *CoRR*, abs/1101.3291, 2011.

[4] M. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of Twitter users. In *SocialCom 2011*, pages 192–199, 2011.

[5] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[7] C. Van Gysel, B. Goethals, and M. de Rijke. Determining the presence of political parties in social circles. In *ICWSM*. AAAI, 2015.

*The original four-page poster was published at the 9th International Conference on Web and Social Media (ICWSM 2015) in May 2015.

Short Text Similarity with Word Embeddings (abstract)

Tom Kenter
tom.kenter@uva.nl

Maarten de Rijke
derijke@uva.nl

University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Determining semantic similarity between two texts is to find out if two pieces of text mean the same thing. Being able to do so successfully is beneficial in many settings in information retrieval like search [7], query suggestion [8], automatic summarization [1] and image finding [3].

In the present work we aim for a generic model that requires no prior knowledge of natural language, such as parse trees, and no external resources of structured semantic information, like Wikipedia or WordNet.

Recent developments in distributional semantics, in particular neural network-based approaches like [9, 11] only require a large amount of unlabelled text data. This data is used to create a, so-called, semantic space. Terms are represented in this semantic space as vectors that are called *word embeddings*. The geometric properties of this space prove to be semantically and syntactically meaningful [4, 9–11], that is, words that are semantically or syntactically similar tend to be close in the semantic space.

A challenge for applying word embeddings to the task of determining semantic similarity of short texts is going from word-level semantics to short-text-level semantics. This problem has been studied extensively over the past few years [2, 6, 12].

In this work we propose to go from word-level to short-text-level semantics by combining insights from methods based on external sources of semantic knowledge with word embeddings. In particular, we perform semantic matching between words in two short texts and use the matched terms to create a saliency-weighted semantic network. A novel feature of our approach is that an arbitrary number of word embedding sets can be incorporated, regardless of the corpus used for training, the underlying algorithm, its parameter settings or the dimensionality of the word vectors. We derive multiple types of meta-features from the comparison of the word vectors for short text pairs and from the vector means of their respective word embeddings, that have not been used before for the task of short text similarity matching.

We show on a publicly available test collection that our generic method, that does not rely on external sources of structural semantic knowledge, outperforms baseline methods that work under the same conditions and outperforms all methods, to our knowledge, that do use external knowledge bases and that have been evaluated on this dataset.

A full version of this paper was presented at CIKM'15 [5].

1. REFERENCES

[1] R. M. Aliguliyev. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications*, 2009.

[2] P. Annesi, D. Croce, and R. Basili. Semantic compositionality in tree kernels. In *CIKM 2014*, 2014.

[3] T. A. Coelho, P. P. Calado, L. V. Souza, B. Ribeiro-Neto, and R. Muntz. Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16 (4):408–417, 2004.

[4] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML 2008*, 2008.

[5] T. Kenter and M. de Rijke. Short text similarity with word embeddings. In *Proceedings of the 24th ACM international conference on information and knowledge management. In CIKM*, volume 15, page 115, 2015.

[6] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[7] H. Li and J. Xu. Semantic matching in search. *Foundations and Trends in Information Retrieval*, 7(5):343–469, 2014.

[8] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. In *ECIR 2007*, 2007.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS 2013*, 2013.

[11] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *EMNLP 2014*, 2014.

[12] R. Socher, E. H. Huang, J. Pennin, C. D. Manning, and A. Y. Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS 2011*, 2011.

(Abstract*) Early Detection of Topical Expertise in Community Question Answering

David van Dijk
d.v.van.dijk@hva.nl

Manos Tsagkias
manos@904labs.com

Maarten de Rijke
derijke@uva.nl

1.[1] TASK

We focus on detecting potential topical experts in community question answering platforms early on in their lifecycle.

2. APPROACH

We extract a set of features (textual, behavioral, time-aware) indicative of a user's expertise on a topic, which we use to train a classifier that learns whether a user shows signals of early expertise given a topic. We cater for early expertise by carefully crafting the training data used to train the classifier.

Although time is a natural way for separating early from seasoned experts, the diverse behavioral patterns among experts make it hard to define early expertise using time in an experimental setting. One future expert might submit ten best answers within two days after joining while another may post one comment during their entire first week. We define expertise based on best answers. Here, a *best answer* is one that gets accepted by the question poster. The more best answers a user gives, the more expert they are. We took as experts those users with one standard deviation number of best answers larger than the average user. On our dataset this translates into people with more than nine best answers on a topic. *Early expertise* is defined as the expertise shown by a user between the moment of joining and becoming an expert, based on the best answers provided. We interpret the values of the selected features between the moment of joining and becoming an expert as the strength of a user's early expertise, and predict future expertise based on it. Table 1 lists the features we used.

Textual features		*Behavioral features*		*Time-aware features*	
1	LM	4	Question	13	Time Interval
2	BM25	5	Answer	14	LM/T
3	TFIDF	6	Comment	15	BM25/T
		7	Z-Score	16	TFIDF/T
		8	Q.-A.	17	Question/T
		9	A.-C.	18	Answer/T
		10	C.-Q.	19	Comment/T
		11	First Answer	20	Z-Score/T
		12	Timely Answer	21	Q.-A./T
				22	A.-C./T
				23	C.-Q./T
				24	First Answer/T
				25	Timely Answer/T

Table 1: The three types of feature: (i) textual, (ii) behavioral, and (iii) time-aware, 25 in total.

3. EXPERIMENTAL SETUP

We want to know the effectiveness of our complete set of features, and of individual feature sets, for classifying users as experts and non-experts; see Table 2 for a summary of systems we consider.

ID	Type	Feature	ID	Feature
A	Textual	1–3	E	C + D
B	Behavioral	4–12	F	A + B
C	Time-aware 1	13–25	G	A + B + C + D
D	Time-aware 2	1–25 per bin		

Table 2: Summary of the systems we consider, and the individual features they consist of.

Our dataset comes from Stack Overflow, and covers the period Aug. 2008–mid-Sept. 2014. We select the 100 most active topic tags in terms of number of questions and answers to maximize the number of experts we can use for training and testing. We mark users as experts on a topic when they have ten or more of their answers marked as best by the question poster, which is one standard deviation larger than the average number of answers over all users and topics.

Machine learning. We start out with unlabeled data and adopt a data-driven approach to labels users who provide above average best answers on a topic as topical experts. Training data for users is generated on the period between joining and becoming an expert. Random Forest was used as classifier for our main experiments.

We report on F1 scores over each best answer of a user starting from their first best answer and going up to their ninth answer, i.e., one best answer before they are deemed experts. At each step, we perform 10-fold cross validation on our test set. We use a two-tailed paired T-test to determine statistical significance and report on significant differences for $\alpha = .05$ and $\alpha = 0.01$.

4. RESULTS

Fig. 1 illustrates the performance of all systems over time. The combination of all features sets (system G) shows the best performance among all systems peaking F1 at 0.7472, and outperforms the baseline (system A) in statistically significant way.

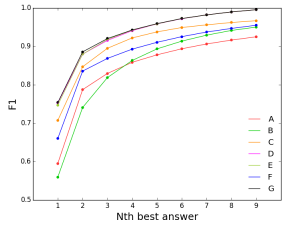

Figure 1: System performance in F1 using individual feature sets and their combination.

We found a small set of features that when combined, perform similarly to our best system (G): TFIDF , Answer, BM25/T, TFIDF/T, Answer/T, Time Interval.

5. CONCLUSIONS

We found that behavioral and temporal features when combined with textual features significantly boost effectiveness peaking F1-score at 0.7472; an 26% improvement over the baseline method. Results demonstrated that our system can accurately predict whether a user will become an expert from a user's first best answer; projected in time, our system makes correct predictions even in 70 months before a user becomes an expert. Although the features to be used may vary, accepted answers are a common phenomenon in QA sites. We therefore expect the method to generalise well within this domain.

[1] D. van Dijk, M. Tsagkias, and M. de Rijke. Early detection of topical expertise in community question answering. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 995–998, 2015.

[1]The full version of this paper was published at SIGIR 2015 [1].

Summarizing Contrastive Themes via Hierarchical Non-Parametric Processes (Abstract)

Zhaochun Ren
z.ren@uva.nl

Maarten de Rijke
derijke@uva.nl

University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Given a topic of interest, a contrastive theme is a group of opposing pairs of viewpoints. We address the task of summarizing contrastive themes: given a set of opinionated documents, select meaningful sentences to represent contrastive themes present in those documents. Several factors make this a challenging problem: unknown numbers of topics, unknown relationships among topics, and the extraction of comparative sentences. Our approach has three core ingredients: contrastive theme modeling, diverse theme extraction, and contrastive theme summarization. Specifically, we present a hierarchical non-parametric model to describe hierarchical relations among topics; this model is used to infer threads of topics as themes from the nested Chinese restaurant process. We enhance the diversity of themes by using structured determinantal point processes for selecting a set of diverse themes with high quality. Finally, we pair contrastive themes and employ an iterative optimization algorithm to select sentences, explicitly considering contrast, relevance, and diversity. Experiments on three datasets demonstrate the effectiveness of our method.

1. INTRODUCTION

Given a set of opinionated documents, we define a *viewpoint* to be a topic with a specific sentiment label, following [5]. A *theme* is a set of viewpoints around a specific set of topics and an explicit sentiment opinion. Given a set of specific topics, two themes are *contrastive* if they are related to the topics, but opposite in terms of sentiment. The phenomenon of contrastive themes is widespread in opinionated web documents [3]. In Fig. 1 we show an example of three contrastive themes about the "Palestine and Israel relationship." Here, each pair of contrastive themes includes two sentences representing two relevant but opposing themes. In this paper, our focus is on developing methods for automatically detecting and describing contrastive themes.

The specific contrastive summarization task that we address in this paper [6] is *contrastive theme summarization of multiple opinionated documents*. In our case, the output consists of contrastive sentence pairs that highlight every contrastive theme in the given documents. To address this task, we employ a non-parametric strategy based on the nested Chinese restaurant process (nCRP) [2]. None of previous work considers the task of contrastive theme summarization. We introduce a topic model that aims to extract contrastive themes and describe hierarchical relations among the underlying topics. Each document in our model is represented by hierarchical threads of topics, whereas a word in each document is assigned a finite mixture of topic paths. We apply collapsed Gibbs sampling to infer approximate posterior distributions of themes.

To enhance the diversity of the contrastive theme modeling, we then proceed as follows. Structured determinantal point processes

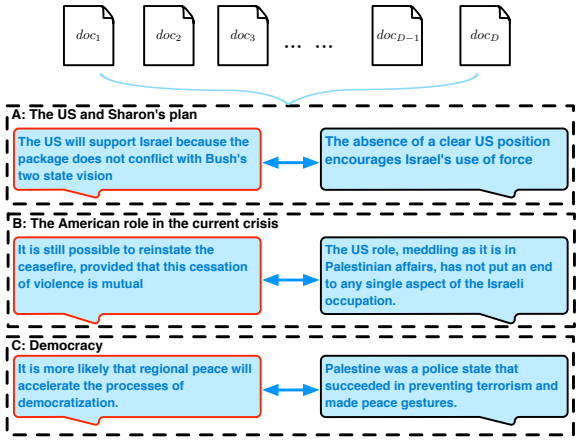

Figure 1: Three example contrastive themes related to "Palestine and Israel." Each contrastive theme shows a pair of opposing sentences.

(SDPPs) [4] are a novel probabilistic strategy to extract diverse and salient threads from large data collections. Given theme distributions obtained via hierarchical sentiment topic modeling, we employ SDPPs to extract a set of diverse and salient themes. Finally, based on themes extracted in the first two steps, we develop an iterative optimization algorithm to generate the final contrastive theme summary. During this process, *relevance*, *diversity* and *contrast* are considered. Our experimental results, obtained using three publicly available opinionated document datasets, show that contrastive themes can be successfully extracted from a given corpus of opinionated documents. Our proposed method for multiple contrastive themes summarization outperforms state-of-the-art baselines, as measured using ROUGE metrics.

2. REFERENCES

[1] A. Ahmed, L. Hong, and A. Smola. Nested chinese restaurant franchise process: Applications to user tracking and document modeling. In *ICML*, 2013.
[2] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2), 2010.
[3] S. Dori-Hacohen and J. Allan. Detecting controversy on the web. In *CIKM*, 2013.
[4] A. Kulesza and B. Taskar. Structured determinantal point processes. In *NIPS*, 2010.
[5] M. J. Paul, C. Zhai, and R. Girju. Summarizing contrastive viewpoints in opinionated text. In *EMNLP*, 2010.
[6] Z. Ren and M. de Rijke. Summarizing contrastive themes via hierarchical non-parametric processes. In *SIGIR*, 2015.

Multileave Gradient Descent for Fast Online Learning to Rank (Abstract)

Anne Schuth[†]
a.g.schuth@uva.nl

Harrie Oosterhuis[†]
harrie.oosterhuis@student.uva.nl

Shimon Whiteson[‡]
shimon.whiteson@cs.ox.ac.uk

Maarten de Rijke[†]
derijke@uva.nl

[†] University of Amsterdam, Amsterdam, The Netherlands
[‡] University of Oxford, Oxford, United Kingdom

1. INTRODUCTION AND METHOD

We summarize the findings of Schuth et al. [6]. Modern search engines base their rankings on combinations of dozens or even hundreds of features. Online learning to rank methods optimize combinations of features while interacting with users of a search engine. Clicks have proven to be a valuable source of information when interpreted as a preference between either rankings [4] or documents [3]. In particular, when clicks are interpreted using interleaved comparison methods, they can reliably infer preferences between a pair of rankers [1, 3]. *Dueling bandit gradient descent* (DBGD) [7] is an online learning to rank algorithm that learns from these interleaved comparisons. It uses the inferred preferences to estimate a gradient, which is followed to find a locally optimal ranker. At every learning step, DBGD estimates this gradient with respect to a *single* exploratory ranker and updates its solution if the exploratory ranker seems better. Exploring *more than one* ranker before updating towards a promising one could lead to finding a better ranker using fewer updates. However, when using interleaved comparisons, this would be too costly, since it would require pairwise comparisons involving users between all exploratory rankers. Instead, we propose to learn from *multileaved comparison methods* [5] that allow for comparisons of multiple rankers at once, using a single user interaction. In this way, our proposed method, *multileave gradient descent* (MGD), aims to speed up online learning to rank. We propose two variants of MGD that differ in how they estimate the gradient. In MGD *winner takes all* (MGD-W), the gradient is estimated using one ranker randomly sampled from those who won the multileaved comparison. In MGD *mean winner* (MGD-M), the gradient is estimated using the mean of all winning rankers. Our contributions are: 1) two approaches, MGD-W and MGD-M, to using multileaved comparison outcomes in an online learning to rank method; and 2) extensive empirical validation of our new methods via experiments on nine learning to rank datasets, showing that MGD-W and MGD-M outperform the state of the art in online learning to rank.

2. RESULTS AND CONCLUSIONS

We run experiments on nine learning to rank datasets and use the setup described by Hofmann et al. [2] to simulate user interactions. Our empirical results, based on extensive experiments encompassing 86M user interactions, show that MGD dramatically improves over the DBGD baseline. In particular, when the noise in user feedback increases, we find that MGD is capable of learning better rankers much faster than the baseline does. Figure 1 shows the results over a larger number of queries. The graph shows that even after 100,000 queries DBGD has not converged, and MGD still performs better.

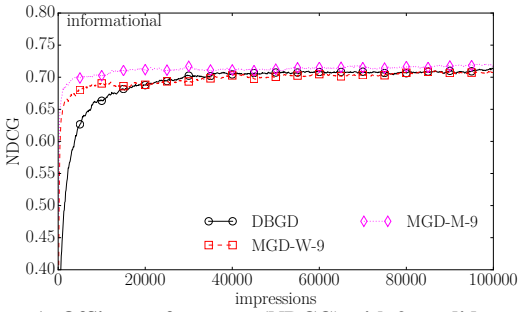

Figure 1: Offline performance (NDCG) with 9 candidates compared to DBGD on *NP2003* dataset.

Generally, after 1,000 query impressions with *noisy* feedback, MGD performs almost on par with DBGD trained on feedback *without any noise*.

An important implication of our results is that orders of magnitude less user interaction data is required to find good rankers when multileaved comparisons are used as feedback mechanism for online learning to rank. This results in far fewer users being exposed to inferior rankers and it allows search engines to adapt faster to changes in user preferences.

REFERENCES

[1] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst.*, 30(1), 2012.

[2] K. Hofmann, A. Schuth, S. Whiteson, and M. de Rijke. Reusing historical interaction data for faster online learning to rank for IR. In *WSDM '13*. ACM, 2013.

[3] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02*. ACM, 2002.

[4] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM '08*. ACM, 2008.

[5] A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved comparisons for fast online evaluation. In *CIKM '14*, pages 71–80. ACM, Nov. 2014.

[6] A. Schuth, H. Oosterhuis, S. Whiteson, and M. de Rijke. Multileave gradient descent for fast online learning to rank. In *WSDM '16*, 2016.

[7] Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *ICML '09*, 2009.

Predicting Relevance based on Assessor Disagreement: Analysis and Practical Applications for Search Evaluation

[Abstract] *

Thomas Demeester[1], Robin Aly[2],
Djoerd Hiemstra[2], Dong Nguyen[2], Chris Develder[1]
[1] Ghent University – iMinds, Belgium
[2] University of Twente, The Netherlands
tdmeeste@intec.ugent.be, r.aly@utwente.nl
{d.hiemstra, d.nguyen}@utwente.nl, cdvelder@intec.ugent.be

ABSTRACT

We present the Predicted Relevance Model (PRM): it allows moving from binary evaluation measures that reflect a single assessor's judgments, towards graded measures that represent the relevance towards random users.

1. INTRODUCTION

Evaluation of search engines relies on assessments of search results for selected test queries, from which we would ideally like to draw conclusions in terms of relevance of the results for general users. In practice, however, most evaluation scenarios only allow us to conclusively determine the relevance towards the particular assessor that provided the judgments. A factor that cannot be ignored when extending conclusions made from assessors towards users, is the possible disagreement on relevance, assuming that a single gold truth label does not exist.

We shortly describe the paper [1] that introduces the Predicted Relevance Model (PRM). The PRM allows predicting a particular result's relevance for a random user, based on an observed assessment and knowledge of the average disagreement between assessors. As a result, existing evaluation metrics designed to measure binary assessor relevance can be transformed into more robust and effectively graded measures that evaluate relevance towards a random user. The PRM also leads to a principled way of quantifying multiple graded or categorical relevance levels for use as gains in established graded relevance measures. Given a single set of test topics with graded relevance judgments, the PRM allows evaluating systems on different scenarios, such as their capability of retrieving top results, or how well they are able to filter out non-relevant ones. Its use in actual evaluation scenarios is illustrated on several information retrieval test collections.

2. THE PREDICTED RELEVANCE MODEL

The following definitions form the core of the PRM: (i) the user population of the search system under evaluation consists of individual users for whom a result is either relevant or non-relevant to a query, (ii) the assessors are part of the evaluation setup, and assign relevance labels to results according to well-described graded (or categorical) assessment levels i, and (iii) the disagreement parameters $p_{R|i}$ represent the probability that a random user would consider a particular result relevant (R), given the knowledge of an independent assessor judgment with level i.

The paper describes these concepts in detail, as well as provides a practical guide for calculating the disagreement parameters based on subsets of double judgments, and an extensive analysis of their properties.

Essential to the PRM are the distinction and the relation between the user model and the assessor model. For example, assessment levels on or above a threshold $i = \theta$ could define a scenario of binary user relevance. As an illustration, consider the task of counting the number N_R of relevant search results among a total set of N results. We denote the number of results assessed with relevance level i as n_i, such that $\sum_i n_i = N$. We can calculate N_R either by neglecting any disagreement between users and assessors (N_R^{bin}), or by taking the disagreement into account (N_R^{PRM}):

$$N_R^{\text{bin}} = \sum_{i \geq \theta} n_i, \quad N_R^{\text{PRM}} = \sum_i n_i \, p_{R|i}. \tag{1}$$

The expression for N_R^{bin} indicates the binary model based on the assessments alone, whereas N_R^{PRM} can be interpreted as the total *expected* number of relevant results for a random user [1]. The difference between both expressions is due to the assessor disagreement, and we show that it cannot be neglected in practice. A similar reasoning leads to the interpretation of metrics such as nDCG from the point of view of a random user, if the gain values are defined by the disagreement parameters, rather than chosen arbitrarily.

In a series of experiments based on existing evaluation collections, it is shown that the PRM leads to a robust evaluation of search engines with respect to several possible notions of binary user relevance.

3. REFERENCES

[1] T. Demeester, R. Aly, D. Hiemstra, D. Nguyen, and C. Develder. Predicting relevance based on assessor disagreement: analysis and practical applications for search evaluation. *Information Retrieval Journal*, 2015.

*A full version [1] of this paper has been accepted (Oct. 2015) for publication in Springer Information Retrieval Journal, special issue on *Information Retrieval Evaluation Using Test Collections*, with DOI: 10.1007/s10791-015-9275-x.

Time-Aware Authorship Attribution of Short Texts

Extended Abstract

Hosein Azarbonyad[1] Mostafa Dehghani[2] Maarten Marx[1] Jaap Kamps[2]

[1]Informatics Institute, University of Amsterdam, The Netherlands
[2]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
{h.azarbonyad,dehghani,maartenmarx,kamps}@uva.nl

ABSTRACT

Automatic authorship attribution is a growing research direction due to its legal and financial importance. In the recent decade with the growth of Internet based communication facilities, much content on the web is in the form of short messages. Finding the author of a short message is important since much fraud and cybercrimes occur with exchanging emails and short messages.

Current authorship attribution approaches neglect an important factor in human development: as a person matures or a significant event occurs in his life (such as changing job, getting married, moving in a new circle of friends, etc) over time the model of his writing style and the words used may change as well. Figure 1 shows the temporal changes of vocabulary usages of 133 Twitter users over a period of 40 months. The figure shows that the similarity of content to a fixed static corpus decreases over time. In fact, we can conclude that content generated at the current time is more similar to recent content than to older content. Current authorship attribution approaches neglect this fact and use all material generated by authors with the same influence.

This paper tries to answer two crucial questions in authorship attribution for short texts. The first research question is: Does the writing style of authors of short text change over time? And if so, do they change their writing styles by the same rate? The second research question is: How does the temporal change of writing styles of authors affect authorship attribution? And how we can capture the changes in the writing styles of authors and take the changes into account to overcome the effects of drift in authorship attribution?

We answer these questions using two datasets: one is collected from Twitter and the Enron email corpus [3]. We introduce a new time-aware authorship attribution approach which is inspired by time-based language models [5] and can be employed in any time-unaware authorship attribution method to consider the temporal drifts in authorship attribution process. We first divide the whole timeline of an author in time periods of a fixed length and then construct a language model for each period. The language model of each period is a probability distribution over n-grams of the texts generated in that period. For a new generated short text, we calculate its similarity with the language model constructed for each period weighted by a decay factor which is a function of the temporal difference of the date of the short text with the period. The time-aware probability that a given short text s is written by an author a is calculated as follows:

$$P(s|a) = \sum_{t \in T \wedge t < t_s} decay(t_s - t) * P(s|\theta_{a_t}), \qquad (1)$$

where T is set of all periods. We discretize the whole timeline to T periods. $P(s|\theta_{a_t})$ is the probability that s is generated by the language model of author a in time period t. The function $decay()$ is a monotonically decreasing function, giving less weight to older periods. We estimate two different decay functions: a general decay function which is same for all authors and a specific decay function for each author estimated based on the change rates of writing styles of authors.

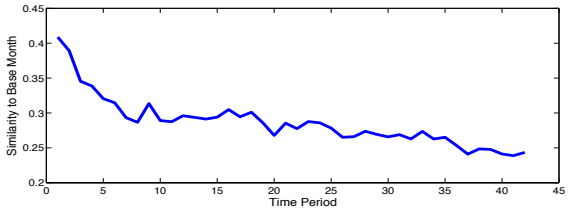

Figure 1: Vocabulary usage changes of Twitter users over time. A dataset containing 133 Twitter users and their written tweets is collected. The first two months of the users' activity in Twitter are considered as start period. Also, each following month is considered as a time period. The $x-axis$ shows the time periods and $y-axis$ shows the averaged similarity of the contents generated by the users at each time period with the content generated by them in the base time period. Cosine similarity over frequency of character 4-grams in users' contents is employed as similarity measure.

We suppose that every author is equally likely before any piece of text is given and finally, the author of s is determined as follows:

$$\hat{a} = \underset{a}{\operatorname{argmax}}\ P(s|a) \qquad (2)$$

We assign s to $\hat{a}$ if $P(s|a)$ is more than a predefined threshold. We use this approach to extend the SCAP method[2] and the feature sampling method [4]. We consider SCAP and feature sampling methods as our baselines. Our evaluations on tweets and Enron datasets show that the proposed time-aware approach is able to incorporate the temporal changes in authors writing styles and outperforms two competitive baselines. The proposed time-aware method improves the accuracy of time-unaware feature sampling baseline by 8% on Enron dataset and by 15% on Tweets dataset. Also this method improves the accuracy of SCAP method by 17% on Enron dataset and by 27% on Tweets dataset.

Acknowledgements The full version of this paper is available as [1]. This research was supported by the Netherlands Organization for Scientific Research(ExPoSe project, NWO CI # 314.99.108; DiLiPaD project, NWO Digging into Data # 600.006.014) and by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement ENVRI, number 283465.

References

[1] H. Azarbonyad, M. Dehghani, M. Marx, and J. Kamps. Time-aware authorship attribution for short text streams. SIGIR '15, pages 727–730, 2015.
[2] G. Frantzeskou, E. Stamatatos, S. Gritzalis, C. E. Chaski, and B. S. Howald. Identifying authorship by byte-level n-grams: The source code author profile (SCAP) method. *IJDE*, 6(1), 2007.
[3] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In *ECML'04*, pages 217–226, 2004.
[4] M. Koppel, J. Schler, and S. Argamon. Authorship attribution in the wild. *LREC*, 45(1):83–94, 2011.
[5] X. Li and W. B. Croft. Time-based language models. CIKM '03, pages 469–475, 2003.

Time-aware Personalized Query Auto Completion

Fei Cai†‡
f.cai@uva.nl

Maarten de Rijke‡
derijke@uva.nl

†Science and Technology on Information Systems Engineering Laboratory,
National University of Defense Technology, Changsha, China
‡University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Query auto completion (QAC) models mostly rank query candidates according to their past popularity. In this paper, we propose a hybrid QAC model that considers two aspects: time-sensitivity and personalization. Using search logs, we return the top N QAC candidates by predicted popularity and rerank them by integrating their similarities with a user's preceding queries. Our experimental results show that our hybrid QAC model outperforms state-of-the-art time-sensitive QAC baseline, achieving around 5% improvements in terms of MRR.

1. INTRODUCTION

Query auto-completion (QAC) takes a few initial keystrokes as input and returns matching queries to auto-complete the search clue. A common approach on QAC is to rank query candidates by their past popularity [1, 3]. But it fails to take strong clues from time, trend and user-specific context into consideration. In our QAC model we first return the top N query completions by predicted popularity, not only based on the recent trend but also based on cyclic phenomena; we then rerank these N completions by user-specific context to output a final query completion list [2].

2. APPROACH

We predict a query q's next-day popularity $\tilde{y}_{t_0+1}(q,\lambda)$ at day t_0+1 before day t_0 by both its recent trend and periodicity with a free parameter λ ($0 \leq \lambda \leq 1$) controlling each contribution:

$$\tilde{y}_{t_0+1}(q,\lambda) = \lambda \times \hat{y}_{t_0+1}(q)_{trend} + (1-\lambda) \times \bar{y}_{t_0+1}(q)_{peri}, \quad (1)$$

where $\lambda = 1$ for aperiodic queries and $0 \leq \lambda < 1$ for periodic queries. The term $\hat{y}_{t_0+1}(q)_{trend}$ is estimated via linear aggregation of predictions from recent N_{days} observations:

$$\hat{y}_{t_0+1}(q)_{trend} = \sum_{i=1}^{N_{days}} norm(\omega_i) \times \hat{y}_{t_0+1}(q,i)_{trend}, \quad (2)$$

where $norm(\omega_i)$ normalizes the contributions from each day to ensure $\sum_i \omega_i = 1$.

The periodicity term $\bar{y}_{t_0+1}(q)_{peri}$ in (1) is smoothed by simply averaging the recent M observations y_{t_p} at preceding time points $t_p = t_0+1-1\cdot T_q, \ldots, t_0+1-M\cdot T_q$ in the log:

$$\hat{y}_{t_0+1}(q)_{peri} = \frac{1}{M}\sum_{m=1}^{M} y_{t_0+1-m\times T_q}(q), \quad (3)$$

where T_q denotes q's periodicity.

Our personalized QAC works here by scoring the candidates $q_c \in \mathcal{S}(p)$ using a combination of similarity scores $Score(Q_s, q_c)$ and $Score(Q_u, q_c)$, where Q_s relates to the recent queries in the current search session and Q_u refers to those of the same user issued before, if available, as:

$$Pscore(q_c) = \omega \cdot Score(Q_s, q_c) + (1-\omega)\cdot Score(Q_u, q_c), \quad (4)$$

where ω controls the weight of the individual component.

Like [1], we then define our hybrid models as convex combinations of two scoring functions:

$$Hscore(q_c) = \gamma \cdot TSscore(q_c) + (1-\gamma)\cdot Pscore(q_c). \quad (5)$$

Next, we use the AOL dataset for evaluation.

3. RESULTS

We present the main results of our model H-QAC in Table 1 compared to those of the state-of-the-art method O-MPC-R. We can see that H-QAC is able to marginally outperform the baseline on both query logs at each prefix length. In contrast with AOL, λ^*-H-QAC on SnV is more sensitive to user's search log with longer prefix. In part, this may be caused by: (1) AOL is a more general search log across topics while SnV focuses on multimedia search. (3) there may be underlying demographic differences between users of the two search logs that lead to changes in query distributions, for example, AOL covers more public users while SnV mostly serves for media professionals.

Table 1: The effectiveness of QAC models on AOL and SnV in terms of MRR, with a query prefix p of 1–5 characters.

	AOL		SnV	
#p	O-MPC-R	H-QAC	O-MPC-R	H-QAC
1	0.1175	**0.1224**$^{\blacktriangle}$	0.2519	**0.2662**$^{\blacktriangle}$
2	0.2027	**0.2091**$^{\blacktriangle}$	0.3607	**0.3907**$^{\blacktriangle}$
3	0.3267	**0.3387**$^{\blacktriangle}$	0.5034	**0.5355**$^{\blacktriangle}$
4	0.4318	**0.4562**$^{\blacktriangle}$	0.6133	**0.6690**$^{\blacktriangle}$
5	0.5087	**0.5236**$^{\blacktriangle}$	0.6992	**0.7491**$^{\blacktriangle}$

4. CONCLUSION

In this paper we extend our time-sensitive QAC method with personalized QAC and verify the effectiveness of our proposal on two datasets, showing significant improvements over various time-sensitive QAC baselines.

REFERENCES

[1] Z. Bar-Yossef and N. Kraus. Context-sensitive query auto-completion. In *WWW '11*, pages 107–116, 2011.

[2] F. Cai, S. Liang, and M. de Rijke. Time-sensitive personalized query auto-completion. In *CIKM '14*, pages 1599–1608, 2014.

[3] M. Shokouhi and K. Radinsky. Time-sensitive query auto-completion. In *SIGIR '12*, pages 601–610, 2012.

DIR'15, November 27, 2015, Amsterdam, The Netherlands.

Learning to Explain Entity Relationships in Knowledge Graphs

[Extended Abstract] *

Nikos Voskarides†
University of Amsterdam
Amsterdam, The Netherlands
n.voskarides@uva.nl

Edgar Meij
Yahoo Labs
London, United Kingdom
emeij@yahoo-inc.com

Manos Tsagkias
904 Labs
Amsterdam, The Netherlands
manos@904labs.com

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

Wouter Weerkamp
904 Labs
Amsterdam, The Netherlands
wouter@904labs.com

ABSTRACT

Knowledge graphs are a powerful tool for supporting a large spectrum of search applications including ranking, recommendation, exploratory search, and web search [3]. A knowledge graph aggregates information around entities across multiple content sources and links these entities together, while at the same time providing entity-specific properties (such as age or employer) and types (such as actor or movie). Although there is a growing interest in automatically constructing knowledge graphs, e.g., from unstructured web data [5, 2], the problem of providing evidence on why two entities are related in a knowledge graph remains largely unaddressed. Extracting and presenting evidence for linking two entities, however, is an important aspect of knowledge graphs, as it can enforce trust between the user and a search engine, which in turn can improve long-term user engagement, e.g., in the context of related entity recommendation [1].

We propose a method for explaining the relationship between two entities, which we evaluate on a newly constructed annotated dataset that we make publicly available.[1] In particular, we consider the task of explaining relationships between pairs of Wikipedia entities. We aim to infer a human-readable description for an entity pair given a relationship between the two entities. Since Wikipedia does not explicitly define relationships between entities we use a knowledge graph to obtain these relations. We cast our task as a sentence ranking problem: we automatically extract sentences from a corpus and rank them according to how well they describe a given relationship between a pair of entities. For ranking purposes, we extract a rich set of features and use learning to rank to effectively combine them. Our feature set includes both traditional information retrieval and natural language processing features that we augment with entity-dependent features. These features leverage information from the structure of the knowledge graph. On top of this, we use features that capture the presence in a sentence of the relationship of interest. For our evaluation we focus on "people" entities and we use a large, manually annotated dataset of sentences.

Our main contributions are a robust and effective method for explaining entity relationships, detailed insights into the performance of our method and features, and a large manually annotated dataset. Our evaluation shows that our method significantly outperforms state-of-the-art sentence retrieval models for this task. Experimental results also show that using relationship-dependent models is beneficial. Our feature analysis shows that relationship type features are the most important, although entity type features are important as well. This indicates that introducing features based on entities identified in the sentences and the relationship is beneficial for this task. Furthermore, the limited dependency on the source feature type indicates that our method might be able to generalize in other domains. Finally, text type features do contribute to retrieval effectiveness, although not significantly.

*The full version of this paper was published at ACL-IJNLP 2015 [4].

†This work was carried out while this author was visiting Yahoo Labs.

[1] `https://github.com/nickvosk/acl2015-dataset-learning-to-explain-entity-relationships`

REFERENCES

[1] R. Blanco, B. B. Cambazoglu, P. Mika, and N. Torzec. Entity recommendations in web search. In *The Semantic Web–ISWC 2013*.

[2] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, (1–2):69–113.

[3] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *KDD 2014*.

[4] N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke, and W. Weerkamp. Learning to explain entity relationships in knowledge graphs. *ACL-IJCNLP 2015*.

[5] J. Weston, A. Bordes, O. Yakhnenko, and N. Usunier. Connecting language and knowledge bases with embedding models for relation extraction. In *EMNLP 2013*.

Lost but Not Forgotten: Finding Pages on the Unarchived Web

Hugo C. Huurdeman[1] Jaap Kamps [1] Thaer Samar[2]
Arjen P. de Vries[2] Anat Ben-David[3] Richard A. Rogers[1]

[1] University of Amsterdam, Amsterdam, the Netherlands, {huurdeman, kamps, r.a.rogers}@uva.nl
[2] Centrum Wiskunde & Informatica, Amsterdam, the Netherlands, {samar, arjen}@cwi.nl
[3] The Open University, Ra'anana, Israel, anatbd@openu.ac.il

ABSTRACT

Web archives attempt to preserve the fast changing web, yet they will always be incomplete. Due to restrictions in crawling depth, crawling frequency, and restrictive selection policies, large parts of the web are unarchived and therefore lost to posterity. In this paper, we propose an approach to uncover unarchived web pages and websites, and to reconstruct different types of descriptions for these pages and sites, based on links and anchor text in the set of crawled pages. We experiment with this approach on the Dutch web archive and evaluate the usefulness of page and host-level representations of unarchived content.

1 Introduction

Every web crawl and web archive is highly incomplete, making the reconstruction of the lost web of crucial importance for the use of web archives and other crawled data. Researchers take the web archive at face value, and equate it to the web as it once was, leading to potentially biased and incorrect conclusions. The main insight of this paper is that although unarchived web pages are lost forever, they are not forgotten in the sense that the crawled pages may contain various evidence of their existence.

We propose a method for deriving representations for unarchived content, by using the evidence of the unarchived web extracted from the collection of archived web pages. We use link evidence to firstly *uncover* target URLs outside the archive, and secondly to *reconstruct* basic representations of target URLs outside the archive. This evidence includes aggregated anchor text, source URLs, assigned classification codes, crawl dates, and other extractable properties. Hence, we derive representations of web pages and websites that are not archived, and which otherwise would have been lost.

2 Unarchived Web Representations

We tested our methods on the data of the selection-based Dutch web archive in 2012. The analysis first characterizes the contents of the Dutch web archive, from which the representations of unarchived pages were subsequently uncovered, reconstructed and evaluated. The archive contains evidence of roughly the same number of unarchived pages as the number of unique pages included in the web archive—a dramatic increase in coverage. In terms of the number of domains and hostnames, the increase of coverage is even more dramatic, but this is partly due to the domain restrictive crawling policy of the Dutch web archive.

However, given that the original page is lost and we rely on indirect evidence, the reconstructed pages have a sparse representation. For a small fraction of popular unarchived pages we have evidence from many links, but the richness of description is highly skewed and tapers off very quickly—we have no more than a few words. This raises doubts on their utility: are these rich enough to distinguish the unique page amongst millions of other pages?

We address this with a critical test cast as a known-item search in a refinding scenario. The evaluation shows that the extraction is rather robust, since both unarchived homepages and non-homepages received similar satisfactory MRR average scores: 0.47 over both types, so on average the relevant unarchived page can be found in the first ranks. Combining page-level evidence into host-level representations of websites leads to richer representations and an increase in retrieval effectiveness (an MRR of 0.63).

3 Discussion and Conclusions

We investigated the recovery of the unarchived pages surrounding the web archive, which we called the 'aura' of the archive. The broad conclusion is that the derived representations are effective, and that we can dramatically increase the coverage of the web archive by our reconstruction approach. This is supported by the fact that only two years since the data was crawled, 20% of the found unarchived homepages and 45% of the non-home pages could no longer be found on the live web nor the Internet Archive.

The unarchived web pages can be used for assessing the completeness of the archive. The recovered pages help to extend the seedlist of the crawlers of selection-based archives, as the pages are potentially relevant to the archive. Additionally, representations of unarchived pages can be used to enrich web archive search systems, and provide additional search functionality. Including the representations of pages in the outer aura, for example, is of special interest as it contains evidence to the existence of top websites that are excluded from archiving, such as Facebook and Twitter.

Acknowledgments This is an extended abstract of [2] and [1]. Funded by NWO (CATCH program, WebART project # 640.005.001).

4 References

[1] H. C. Huurdeman, A. Ben-David, J. Kamps, T. Samar, and A. P. de Vries. Finding pages on the unarchived web. In *IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom, September 8-12, 2014*, pages 331–340. IEEE, 2014. http://dx.doi.org/10.1109/JCDL.2014.6970188.

[2] H. C. Huurdeman, J. Kamps, T. Samar, A. P. de Vries, A. Ben-David, and R. A. Rogers. Lost but not forgotten: finding pages on the unarchived web. *Int. J. on Digital Libraries*, 16:247–265, 2015. http://dx.doi.org/10.1007/s00799-015-0153-3.

DIR'15 November 27, 2015, Amsterdam, The Netherlands.

Behavioral Dynamics from the SERP's Perspective: What are Failed SERPs and How to Fix Them?

Julia Kiseleva[1] Jaap Kamps[2] Vadim Nikulin[3] Nikita Makarov[3]

[1]Eindhoven University of Technology, Eindhoven, The Netherlands, j.kiseleva@tue.nl
[2]University of Amsterdam, Amsterdam, The Netherlands, kamps@uva.nl
[3]Yandex, Moscow, Russian Federation, {vnik,nkmakarov}@yandex-team.ru

ABSTRACT

Web search is always in a state of flux: queries, their intent, and the most relevant content are changing over time, in predictable and unpredictable ways. Modern search technology has made great strides in keeping up to pace with these changes, but there remain cases of failure where the organic search results on the search engine result page (SERP) are outdated, and no relevant result is displayed. Failing SERPs due to temporal drift are one of the greatest frustrations of web searchers, leading to search abandonment or even search engine switch. Detecting failed SERPs timely and providing access to the desired out-of-SERP results has huge potential to improve user satisfaction.

1. RESULTS

Our main research question was: By analyzing behavioral dynamics at the SERP level, can we detect an important class of detrimental cases (such as search failure) based on changes in observable behavior caused by low user satisfaction?We presented an overview of prior work on topic and concept drift, behavioral dynamics, and user satisfaction on the web, with a special focus on the $SERP$ level. We conducted a conceptual analysis of success and failure at the SERP level in order to answer our first research question: How to include the SERP into the conceptual model of behavioral dynamics on the web? How to identify (un)-successful SERPs in terms of drastic changes in observable user behavior? Specifically, we introduced the concept of a successful and failed SERP and analyzed their behavioral consequences identifying indicators of success and failure. By analyzing success and failure in light of changing query intents over time, we identified an important case of SERP failure due to query intent drift. This suggested an approach to detect a failed SERP due to query intent drift by significant changes in behavioral indicators of failure.

We continued our analysis of different types of drifts in query intent over time, answering our second research question: Can we distinguish different types of SERP failure due to query intent drift (e.g., sudden, incremental), and when and how should we update the SERP to reflect these changes? Inspired by the literature on concept drift [1], we studied different changes in query intent: sudden, incremental, gradual and reoccurring, and identified relevant parameters, such as the window of change, volume or popularity of queries, and relevant behavioral indicators, such as the probability of reformulation, abandonment rates, and click through rates. For the two main categories of intent drift, we define an unsupervised approach to detect failed SERPs caused by drift, requiring only a single pass through a transaction log.

Finally, we ran experiments on massive raw search logs, answering our third research question: How effective is our approach on a realistic sample of traffic of a major internet search engine? We ran a simplified version of our algorithm and detected over 200,000 pairs of $\langle Q, SERP \rangle$ suspected of failing due to drifting query intents, observing a reasonable accuracy of drift detection (72%) and a high accuracy of candidate URLs to be included on the SERP of the original query. For incremental change over the longer detection period of 14 days, we detected failed SERPs due to query intent drift with an 80% accuracy, but under the specific conditions of the recency optimized search engine the performance for detecting sudden change over shorter periods was less effective.

Our analysis of behavioral dynamics at the SERP level gives new insight in one of the primary causes of search failure due to temporal query intent drifts. Our overall conclusion is that the most detrimental cases in terms of (lack of) user satisfaction lead to the largest changes in information seeking behavior, and hence to observable changes in behavior we can exploit to detect failure, and moreover not only detect them but also resolve them.

Acknowledgments This is an extended abstract of [2].

2. REFERENCES

[1] J. Gama, I. Zliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014.

[2] J. Kiseleva, J. Kamps, V. Nikulin, and N. Makarov. Behavioral dynamics from the serp's perspective: What are failed serps and how to fix them? In J. Bailey, A. Moffat, C. C. Aggarwal, M. de Rijke, R. Kumar, V. Murdock, T. Sellis, and J. X. Yu, editors, *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1561–1570. ACM, 2015.

DIR'15 November 27, 2015, Amsterdam, The Netherlands.

Categorizing Events using Spatio-Temporal and User Features from Flickr

[Abstract] *

Steven Van Canneyt[1], Steven Schockaert[2], Bart Dhoedt[1]
[1] Ghent University – iMinds, Belgium
[2] Cardiff University, United Kingdom
steven.vancanneyt@ugent.be, schockaerts1@cardiff.ac.uk
bart.dhoedt@ugent.be

ABSTRACT

We introduce a method for discovering the semantic type of events extracted from Flickr, focusing in particular on how this type is influenced by the spatio-temporal grounding of the event, the profile of its attendees, and the semantic type of the venue and other entities which are associated with the event.

1. INTRODUCTION

Even though the problem of event detection from social media has been well studied in recent years, few authors have looked at deriving structured representations for their detected events. We envision the use of social media for extracting large-scale structured event databases, which could in turn be used for answering complex (historical) queries. In this paper, we study how the semantic type of events can be extracted from social media, as a first step towards automatically extending and creating structured event databases.

2. METHODOLOGY

Evidence about the semantic type of an event can be obtained by analyzing social media documents, such as Flickr photos taken at the event, which we consider in this paper, or tweets that have been sent about the event. In particular, we represent an event as a set of social media documents related to that event, together with its associated characteristics. A set of social media documents related to an event may for instance be automatically extracted from social media or may be extracted from existing event databases such as Upcoming. Most initial work about discovering the semantic types of events only used textual information, which may lead to poor performance when the text is noisy (e.g. in some Twitter posts) or absent (e.g. in some Flickr photos). However, social media documents also contain metadata which provide an indication about the spatio-temporal and attendees features of an event. The hypothesis we consider in this paper is that in many cases the event type can be discovered by looking at characteristics, such as timing, the location of the event or properties of attendees, which can be readily obtained from social media sources. For example, when an event occurs on a Saturday inside a sport complex and it has basketball players as main actors, it is very likely that this event is of type 'basketball game'.

Even though our methods can be applied more generally, we restrict ourselves in this paper to experiments with Flickr photos. In particular, the spatio-temporal grounding of the event, the profile of its attendees, and the semantic type of the venue and other entities which are associated with the event are estimated using its associated Flickr photos, and these characteristics are then used to describe the event. To estimate the type of a given event, we use an ensemble of classifiers, one for each of the considered descriptors.

3. EVALUATION

We consider two use cases. First, the trained classifiers are used to analyze in detail to what extent our methodology is able to discover the semantic type of known events that have no associated semantic type. This is useful, for example, to improve existing event databases such as Upcoming, for which we found that about 10% had no known type. When using our methodology instead of a baseline which only uses the text of the Flickr photos related to an event to estimate its semantic type, the classification accuracy increases significantly. We observe that considering the type of the events visited in the past by the participants of the event lead to the most substantial improvement over the baseline approach. The classification performance is further improved when the types of known events organized nearby the event, the textual content of the photos taken in the vicinity of the event, and the time and date of the event are considered. Second, the model is used to estimate the semantic type of events which have been automatically detected from Flickr and are not mentioned in existing event datasets, which could substantially increase the applicability of existing methods for automated event detection.

4. ACKNOWLEDGMENTS

Steven Van Canneyt is funded by a Ph.D. grant of the Agency for Innovation by Science and Technology (IWT).

5. REFERENCES

[1] S. Van Canneyt, S. Schockaert, and B. Dhoedt. Categorizing events using spatio-temporal and user features from flickr. *Information Sciences*, 328:76–96, 2016.

*A full version [1] of this paper has been accepted and will be published (Jan. 2016) in Information Sciences.

CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data

Anca Dumitrache
Vrije Universiteit Amsterdam
anca.dumitrache@vu.nl

Oana Inel
Vrije Universiteit Amsterdam
oana.inel@vu.nl

Benjamin Timmermans
Vrije Universiteit Amsterdam
b.timmermans@vu.nl

Lora Aroyo
Vrije Universiteit Amsterdam
lora.aroyo@vu.nl

Robert-Jan Sips
IBM CAS Netherlands
robert-jan.sips@nl.ibm.com

Keywords
crowdsourcing, gold-standard, machine-human computation, data analysis, experiment replication

1. ABSTRACT
Information Retrieval (IR) systems typically use for training and evaluation gold standard annotations, i.e *ground truth*. Traditionally ground truth is collected by asking domain experts to annotate a number of examples and by providing them with a set of annotation guidelines to ensure an uniform understanding of the annotation task. This process is entirely based on a simplified notion of truth, i.e. under the assumption that a single right annotation exists for each example. However, in reality we continuously observe that truth is not universal and is strongly influenced by the variety of factors, e.g. context, background knowledge, points of view, as well as the quality of the examples themselves.

Research in IR has started to incorporate crowdsourcing in designing, training and evaluating information retrieval systems [4]. Using crowdsourcing platforms such as CrowdFlower or Amazon Mechanical Turk for gathering human interpretation on data has become now a mainstream process. However, as we have observed previously [1], the introduction of crowdsourcing has not fundamentally changed the way gold standards are created: humans are still asked to provide a semantic interpretation of data, with the explicit assumption that there is *one correct interpretation*. Thus, the diversity of interpretation and perspectives is still not taken into consideration - neither in the training, nor in the evaluation of such systems.

In previous work, we introduced the *CrowdTruth methodology*, a *novel approach for gathering annotated data from the crowd*. Inspired by the simple intuition that human interpretation is subjective [1], and by the observation that disagreement is a natural product of having multiple people performing annotation tasks, CrowdTruth can provide useful insights about the task design, annotation clarity, or annotator quality. We reject the traditional notion of ground truth in gold standard annotation, in which annotation tasks are viewed as having a single correct answer, and adopt instead a disagreement-based ground truth, we call *CrowdTruth*. In previous experiments [3, 2] we have validated the *CrowdTruth* metrics for example, worker and target annotation quality in a variety of annotation tasks, data modalities and domains. We showed experimental evidence that these metrics are interdependent, and that measuring only worker quality is missing an important information, as not all the annotated units are created equal.

In this paper, we introduce the CrowdTruth open-source machine-human computing framework that implements the *CrowdTruth Methodology* for gathering annotations on different types of data and in different domains[1,2,3,4]. We combine in an optimized workflow the best of both worlds, i.e. human accuracy in semantic interpretation and machine abilities to process massive amounts of data.

The main concept behind the CrowdTruth methodology is employing a comparatively large number of crowd annotators per unit. Inter-annotator disagreement is then modeled using CrowdTruth metrics based on cosine similarity. Our work in medical relation extraction [3] has shown that this methodology can help us find evidence of ambiguous sentences that are difficult to classify. Considering the growing number of crowdsourcing usage in the IR community, and the growing need for gold standard data, we believe CrowdTruth can be of critical relevance to provide a scientific methodology for using crowdsourcing in a reliable and replicable way.

2. REFERENCES
[1] L. Aroyo and C. Welty. Truth Is a Lie: CrowdTruth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24, 2015.

[2] V. de Boer et al. Dive in the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on WWW*, 2015.

[3] A. Dumitrache, L. Aroyo, and C. Welty. CrowdTruth Measures for Language Ambiguity. In *Proc. of LD4IE Workshop, ISWC*, 2015.

[4] M. Lease and E. Yilmaz. Crowdsourcing for information retrieval: introduction to the special issue. *Information retrieval*, 16(2):91–100, 2013.

[1] framework: `https://github.com/CrowdTruth/CrowdTruth`
[2] service: `http://CrowdTruth.org`
[3] documentation: `http://CrowdTruth.org/info`
[4] datasets: `http://data.CrowdTruth.org`

Dynamic Collective Entity Representations for Entity Ranking (Abstract)

David Graus
University of Amsterdam
d.p.graus@uva.nl

Manos Tsagkias
904Labs
manos@904labs.com

Wouter Weerkamp
904Labs
wouter@904labs.com

Edgar Meij
Yahoo Labs
emeij@yahoo-inc.com

Maarten de Rijke
University of Amsterdam
derijke@uva.nl

1. INTRODUCTION AND METHOD

We summarize the findings of Graus et al. [2]. Many queries issued to search engines are related to entities [4]. Entity ranking, where the goal is to position a relevant entity at the top of the ranking for a given query, is therefore becoming an ever more important task [1]. Entity ranking is inherently difficult due to the potential mismatch between the entity's description in a knowledge base and the way people refer to the same entity when searching for it.

We propose a method that aims to close this gap by leveraging collective intelligence as offered by external entity "description sources". We differentiate between dynamic description sources that are timestamped, and static ones that are not. We leverage five static description sources for expanding entity representations. First, from the knowledge base: (1) anchor text of inter-knowledge base hyperlinks, (2) redirects, (3) category titles, and (4) names of entities that are linked to or from an entity. From the web: (5) web anchors that link to entities. In addition, we leverage three dynamic description sources, whose content are added to entity representations in a streaming manner: (6) search engine queries that yield clicks on entities, (7) tweets, and (8) tags that mention entities.

We represent entities as fielded documents [5], where each field corresponds to content that comes from one description source. As external description sources continually come in, the content in the entity's fields changes, and previously learned feature weights may be sub-optimal. Hence, constructing a dynamic entity representation for optimal retrieval effectiveness boils down to dynamically learning to optimally weight the entity's fields. We exploit implicit user feedback (i.e., clicks) to retrain our model, and relearn the weights associated to the fields, much like online learning to rank [3]. As an entity ranker, we employ a random forest classifier, using its confidence scores as a ranking signal.

2. RESULTS AND CONCLUSIONS

To evaluate our method's performance over time, we treat users' clicks as ground truth, and the goal is to rank clicked entities at position 1. We split the query log into chunks, allocate the first chunk for training, and evaluate each succeeding query in the next chunk. Then, we add this chunk to the training set, retrain the classifier, and continue evaluating the next chunk. In Figure 1 we compare our Dynamic Collective Entity Representation method (`DCER`) to a static baseline, which only exploits the Knowledge Base description sources (`KBER`), i.e., sources 1–4 in Section 1. We see how each individual description source contributes to more effective ranking, with `KB+tags` narrowly outperforming `KB+web` as the best single source. We observe that after about 18,000 queries, `KB+tags` overtakes the (static) `KB+web` method, suggesting that newly incoming tags yield higher ranking effectiveness.

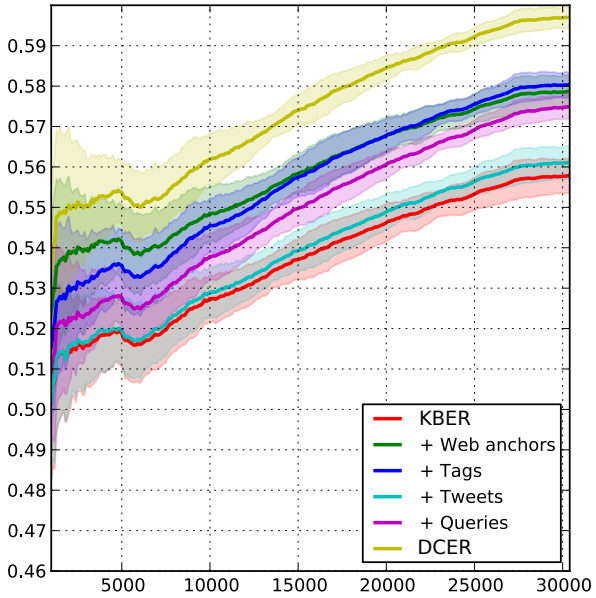

Figure 1: Impact on performance of individual description sources. MAP on the y-axis, number of queries on the x-axis. This plot is best viewed in color.

Our results demonstrate that incorporating dynamic description sources into dynamic collective entity representations enables a better matching of users' queries to entities. Furthermore, we show how continuously updating the ranker leads to improved ranking effectiveness in dynamic collective entity representations.

REFERENCES

[1] K. Balog, P. Serdyukov, and A. de Vries. Overview of the TREC 2010 entity track. In *TREC 2010*, 2011.

[2] D. Graus, M. Tsagkias, W. Weerkamp, E. Meij, and M. de Rijke. Dynamic collective entity representations for entity ranking. In *WSDM '16*, 2016.

[3] K. Hofmann, S. Whiteson, and M. de Rijke. A probabilistic method for inferring preferences from clicks. In *CIKM '11*, pages 249–258, 2011.

[4] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *WWW '10*, pages 561–570, 2010.

[5] C. Macdonald, R. L. Santos, I. Ounis, and B. He. About learning models with multiple query-dependent features. *ACM TOIS*, 31(3):11:1–11:39, 2013.

Mining, Ranking and Recommending Entity Aspects (Abstract)

Ridho Reinanda[†]
r.reinanda@uva.nl

Edgar Meij[‡]
emeij@yahoo-inc.com

Maarten de Rijke[†]
derijke@uva.nl

[†] University of Amsterdam, Amsterdam, The Netherlands
[‡] Yahoo Labs, London, United Kingdom

ABSTRACT

With the proliferation of mobile devices, an increasing amount of available structured data, and the development of advanced search result pages, modern-day web search is increasingly geared towards entity-oriented search [1, 7, 9]. A first step and common strategy to address such information needs is to identify entities within queries, commonly known as *entity linking* [6]. Semantic information that is gleaned from the linked entities (such as entity types, attributes, or related entities) is used in various ways by modern search engines, e.g., for presenting an entity card, showing actionable links, and/or recommending related entities [2, 4, 5].

Entities are not typically searched for on their own, however, but often combined with other entities, types, attributes/properties, relationships, or keywords [9]. Such query completions in the context of an entity are commonly referred to as entity-oriented intents or entity aspects [8, 11]. In this paper we study the problem of mining and ranking entity aspects in the context of web search. In particular, we study four related tasks in this paper: (1) identifying entity aspects, (2) estimating the importance of aspects with respect to an entity, (3) ranking entity aspects with respect to a current query and/or user session, and (4) leveraging entity aspects for query recommendation.

The first step in identifying entity aspects involves extracting common queries in the context of an entity and grouping them based on their similarity. We perform this process offline and investigate three matching strategies for clustering queries into entity aspects: *lexical*, *semantic*, and *click-based*. Gathering such entity aspects can already be valuable on its own since they can be used to, e.g., discover bursty or consistent entity intents or to determine entity type-specific aspects [8].

In the next step we rank the obtained entity aspects for each entity in a query-independent fashion using three distinct strategies. This provides us with a mechanism to retrieve the most relevant aspects for a given entity on its own, which, in turn, can be used to, e.g., summarize the most pertinent information needs around an entity or to help the presentation of entity-oriented search results such as customized entity cards on SERPs [1].

The third task that we consider is aspect recommendation. Given an entity and a certain aspect as input, recommend related aspects. This task is motivated by the increasing proliferation of entity-oriented interface elements for web search that can be improved by, e.g., (re)ranking particular items on these elements. Recommending aspects for an entity can also help users discover new and serendipitous information with respect to an entity. We consider two approaches to recommend aspects: *semantic* and *behavioral*. In the semantic approach, relatedness is estimated from a semantic representation of aspects. The behavioral approach is based on the "flow" of aspect transitions in actual user sessions, modeled using an adapted version of the query-flow graph [3].

We perform large-scale experiments on both a publicly available and a commercial search engine's query log to evaluate our proposed methods for mining, ranking, and recommending entity aspects, as well as for recommending queries. We perform contrastive experiments using various similarity measures and ranking strategies. We evaluate the quality of the extracted entity aspects by manually evaluating the generated clusters. Since manually evaluating aspect rankings for entities is not straightforward, we resort to automatic evaluation based on adjacent query pairs. A similar automatic evaluation strategy is also employed to evaluate aspect recommendations.

We find that entropy-based methods achieve the best performance compared to maximum likelihood and language modeling on the task of entity aspect ranking. Concerning aspect recommendation we find that combining aspect transitions within a session and semantic relatedness give the best performance. Furthermore, we show that the entity aspects can be effectively utilized for query recommendation.

This work was presented at SIGIR 2015 [10].

References

[1] N. Balasubramanian and S. Cucerzan. Topic pages: An alternative to the ten blue links. In *IEEE-ICSC 2010*, 2010.
[2] R. Blanco, B. B. Cambazoglu, P. Mika, and N. Torzec. Entity recommendations in web search. In *ISWC '13*, 2013.
[3] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: Model and applications. In *CIKM '08*, 2008.
[4] L. Hollink, P. Mika, and R. Blanco. Web usage mining with semantic analysis. In *WWW '13*, 2013.
[5] T. Lin, P. Pantel, M. Gamon, A. Kannan, and A. Fuxman. Active objects: Actions for entity-centric search. In *WWW '12*, 2012.
[6] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM 2012*, 2012.
[7] P. Pantel and A. Fuxman. Jigs and lures: Associating web queries with structured entities. In *ACL '11*, 2011.
[8] P. Pantel, T. Lin, and M. Gamon. Mining entity types from query logs via user intent modeling. In *ACL'12*, 2012.
[9] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *WWW '10*, pages 771–780, 2010.
[10] R. Reinanda, E. Meij, and M. de Rijke. Mining, ranking and recommending entity aspects. In *SIGIR 2015: 38th international ACM SIGIR conference on Research and development in information retrieval*. ACM, August 2015.
[11] X. Yin and S. Shah. Building taxonomy of web search intents for name entity queries. In *WWW '10*, 2010.

Eye-tracking Studies of Query Intent and Reformulation

Carsten Eickhoff
Dept. of Computer Science
ETH Zurich, Switzerland
ecarsten@inf.ethz.ch

Sebastian Dungs
University of Duisburg-Essen
Duisburg, Germany
dungs@is.inf.uni-due.de

Vu Tran
University of Duisburg-Essen
Duisburg, Germany
vtran@is.inf.uni-due.de

Users of information retrieval systems have been frequently shown to struggle with forming accurate mental images of their information needs and the resources to satisfy them. Belkin et al. describe this deficit as an *Anomalous State of Knowledge* (ASK), hindering users' query formulation and search success [1]. To mitigate this effect, Web search engines offer query suggestions and recommendations that guide the searcher towards popular queries, frequently issued by other users. It is, however, often unclear how relevant such suggestions are for the individual user, especially for non-navigational information needs. Ideally, we would like to promote those suggestions, that lead the user to relevant, novel and understandable documents rather than just generally popular ones. Personalized generation of query suggestions based on the user's previous search and interaction history has been found as one way to address this problem [3]. The proposed models, however, are coarse-grained and represent only high-level notions of the user's active query vocabulary. They consider, for example, all previously encountered terms (*e.g.*, all terms present on recently visited Web sites) to be known and understandable. While this family of approaches makes a valuable first step towards integrating an understanding of the user's state of knowledge into the query suggestion process, one would require a system that can account for the user's vocabulary at a significantly finer granularity, ideally on the term level.

The same issue plays up at other points of the search process, for example during result ranking. State-of-the-art relevance models often include representations of the user and their specific context such as previous search history [8], preferences in terms of high-level topics [7], or content readability [4]. While such notions of text complexity have been demonstrated to significantly increase retrieval performance by providing users with resources of appropriate reading level, the readability metrics themselves are not personalized and rely on general complexity estimates based on a very diverse audience of users. Instead, it would be strongly desirable to know which exact terms the searcher is able to recognize, understand and actively use.

In this paper, we use eye-gaze fixations and cursor movement information in order to study which concrete terms the user pays most attention to on Web pages and search engine result pages (SERPs) and how those terms are subsequently re-used as query terms. Our experiments suggest that terms receiving the most user attention in terms of gaze frequency and fixation duration are significantly more likely to be used as future query terms. In contrast to previous assumptions from log-based studies [6], we show that only as few as 21% of all newly added query terms within a session can be explained as verbatim copies of previously seen terms. To investigate the origin of the remaining terms, we turn away from our study of literal term matches and consider semantic relatedness of fixated and newly added terms, instead. As for literal matches before, we note that prospective query terms are semantically highly related to those terms that were previously fixated on. In a series of dedicated follow-up experiments, we account for the effects of individual term informativeness, frequency, length and complexity that could potentially interfere with our observations. We note that while long, rare and more complex terms receive more attention than their more frequent and general counterparts, user interest remains the governing factor. Encouraged by this observation, we turn towards a practical application of the thus measured user intent in a query suggestion setting. Re-ranking query suggestion candidates according to term-level estimates of user intent results in significantly improved performance as compared to the raw output of a popular search engine as well as state-of-the-art paragraph-level intent models [2]. In order to allow for broad exploitation, a final series of experiments confirms the robustness of our findings when replacing eye-gaze traces with more widely accessible cursor movement information.

This work first appeared as a SIGIR 2015 full paper [5].

1. REFERENCES

[1] N. Belkin, R. Oddy, and H. Brooks. Ask for information retrieval: Part i. background and theory. *Journal of documentation*, 1982.

[2] G. Buscher, A. Dengel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. In *SIGIR 2008*. ACM.

[3] P. Chirita, C. Firan, and W. Nejdl. Personalized query expansion for the web. In *SIGIR 2007*. ACM.

[4] K. Collins-Thompson, P. Bennett, R. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *CIKM 2011*. ACM.

[5] C. Eickhoff, S. Dungs, and V. Tran. An eye-tracking study of query reformulation. In *SIGIR 2015*.

[6] C. Eickhoff, J. Teevan, R. White, and S. Dumais. Lessons from the journey: A query log analysis of within-session learning. In *WSDM 2014*. ACM.

[7] D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. In *JCDL 2002*. ACM.

[8] J. Teevan, S. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR 2005*. ACM.

From Multistage Information-Seeking Models to Multistage Search Systems

Hugo C. Huurdeman[1], Jaap Kamps[1]

[1] University of Amsterdam, Amsterdam, The Netherlands, {huurdeman,kamps}@uva.nl

ABSTRACT

This extended abstract summarizes research from [1].

The ever expanding digital information universe makes us rely on search systems to sift through immense amounts of data to satisfy our information needs. Our searches using these systems range from simple lookups to complex and multifaceted explorations. A multitude of models of the information seeking process, for example Kuhlthau's ISP model, divide the search process for complex search tasks into multiple stages. Current search systems, in contrast, still predominantly use a "one-size-fits-all" approach: one interface is used for all stages of a search, even for complex search endeavors. The main aim of this paper is to bridge the gap between multistage information seeking models, documenting the search process on a general level, and search systems and interfaces, serving as the concrete tools to perform searches.

1. Multistage Information Seeking Models Various temporally-based information seeking models differentiate search stages over time, based on empirical evidence. During these stages, the information sought, the relevance, and the search tactics and strategies evolve. Authors like Kuhlthau and Vakkari have accurately pinpointed the issue of stage-specific search support, but provide less concrete pointers to implementation in search systems and interfaces. Many information seeking models, as Tom Wilson has indicated, focus on higher-level aspects of information seeking (the *macro* level), while information system designers usually focus on the support for concrete actions of a searcher (the *micro* level). However, indications for the provision of search stage support in search systems can be determined from the theory, not only at the interface level (providing specific features supporting stages), but also at the system level (for example providing search stage adaptive ranking). Our main conclusion from this conceptual analysis is that a good general understanding of the information seeking stages exists at the macro level, but that the translation into system and user interface design choices at the micro level remains unsolved.

2. SUI Support for Information Seeking To get more insights into the Search User Interface (SUI) features that could support complex, information-intensive search tasks, the second part of the paper analyzes concrete SUI features using Max Wilson's framework for interface features, which differentiates between *input*, *control*, *informational* and *personalizable* features. We argue that there is an abundance of interfaces which support information *search*, but that few systems provide explicit support for the higher-level information *seeking* process in the context of complex tasks. However, some overarching interface paradigms have similarities with temporal search stages. *Exploratory search*, though slightly different in nature due to the open-endedness of the tasks, could fit in the early 'prefocus' stages of Kuhlthau's and Vakkari's models, and elements of *sensemaking* could fit in the more advanced 'postfocus' stages of search. There is, however, no integrated system, and various authors point at the complexity to understand the impact of design choices on the overall usability, and the complexity of creating a seamless and effortless flow of interaction. This part of the paper concludes that there is a good understanding of search user interface features at the micro level, but that our general understanding of behavior at the macro level is fragmented at best. This suggests a way to reconcile these two views: what if we utilize the understanding of information seeking models at the macro level as a way to understand the flow of interaction at the micro level?

3. Interface Features and Search Stage In this section, we analyze the influence of search stage on the flow of interaction. We observe different use of features over time, based on previous literature and an analysis of eye tracking and system data from a small-scale user study. Some *informational* features (results lists and details) are generally used in all stages of the search, albeit in different depths, and therefore could be considered stage insensitive. However, the use of a subset of search features varied over time, like the gaze towards the query box (an *input* feature), and the use of the basket (a *personalizable* feature). Especially, we observe variations in the use of interface features in the beginning and end of a complex search task. This provides indications of different usage patterns of search user interface features in different search stages, which could be informative for the design of search systems.

4. Conclusion The main contribution of this paper is that it conceptually reconciles macro level information seeking stages and micro level search system features. We highlight the impact of search stages on the flow of interaction with SUI features, providing new handles for the design of multistage search systems. Based on our analysis of information seeking models, search user interfaces and search feature use over time, we hypothesize that there are differences in the interaction flow of SUI feature use at the micro level, depending on the current stage of search at the macro level. This suggests interface elements which are search stage sensitive, and we could customize the way search system features are shown during task progression. This customization may be performed in different ways: depending on the search stage, one could adaptively show SUI features, adjust the shown details of features, or change their prominence, position and size. In follow-up research we investigate whether this approach can be naturally integrated in the user's flow, for different complex tasks and contexts.

Acknowledgements The authors wish to thank Vu Tran (Univ. of Duisburg-Essen). Funded by NWO (CATCH, # 640.005.001).

REFERENCES

[1] H. C. Huurdeman and J. Kamps. From Multistage Information-seeking Models to Multistage Search Systems. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX)*, IIiX '14, pages 145–154, 2014. ACM. http://dx.doi.org/10.1145/2637002.2637020.

DIR'15 November 27, 2015, Amsterdam, The Netherlands.

Struggling and Success in Web Search (Abstract) *

Daan Odijk
University of Amsterdam
Amsterdam, The Netherlands
d.odijk@uva.nl

Ryen W. White, Ahmed Hassan Awadallah, Susan T. Dumais
Microsoft Research, Redmond, WA, USA
{ryenw,hassanam,sdumais}@microsoft.com

ABSTRACT

Web searchers sometimes struggle to find relevant information. Struggling leads to frustrating and dissatisfying search experiences, even if searchers ultimately meet their search objectives. When searchers experience difficulty in finding information, their struggle may be apparent in search behaviors such as issuing numerous search queries or visiting many results within a search session. Such long sessions are prevalent and time consuming (e.g., around half of Web search sessions contain multiple queries). Long sessions occur when searchers are exploring or learning a new area, or when they are struggling to find relevant information. Methods have recently been developed to distinguish between struggling and exploring in long sessions using only behavioral signals [1]. However, little attention has been paid to *how* and *why* searchers struggle. This is particularly important since struggling is prevalent in long sessions, e.g., Hassan et al. [1] found that in 60% of long sessions, searchers' actions suggested that they were struggling. Better understanding of struggling search sessions is important in improving search systems.

Fig. 1 shows an example struggling session of a searcher interested in watching live streaming video of the U.S. Open golf tournament.

9:13:11 AM **Query** us open
9:13:24 AM **Query** us open golf
9:13:36 AM **Query** us open golf 2013 live
9:13:59 AM **Query** watch us open live streaming
9:14:02 AM **Click** http://inquisitr.com/1300340/watch-2014-u-s-open-live-online-final-round-free-streaming-video
9:31:55 AM **END**

Figure 1: Example of a *struggling* session from June 2014.

The first two queries yield generic results about U.S. Open sporting events and the specific tournament. The third query might have provided the correct results but it included the previous year rather than the current year. At this stage, the searcher appears to be struggling. The fourth query is the so-called *pivotal query* where the searcher drops the year and adds the terms "watch" and "streaming". This decision to add these terms alters the course of the search session and leads to a seemingly successful outcome.

Understanding transitions between queries in a struggling session, and transitions between struggling and successful queries, can inform the development of strategies and algorithms to help reduce struggling. We address this issue using a mixed methods study using large-scale logs, crowd-sourced labeling, and predictive modeling.

We analyze anonymized search logs from the Microsoft Bing Web search engine to characterize aspects of struggling searches and to understand how some sessions result in success, while others result in failure. Through log analysis on millions of these search sessions, we show that there are significant differences in how struggling searchers behave given different outcomes. These differences encompass many aspects of the search process, including queries, query reformulations, result click behavior, landing page dwell time, and the nature of the search topic. We find that struggling searchers issue fewer queries in successful sessions than in unsuccessful ones. In addition, queries in unsuccessful sessions are shorter, fewer results are clicked and the query reformulations indicate that searchers have more trouble choosing the correct vocabulary.

Given these intriguing differences, we employ a crowd-sourcing methodology to broaden our understanding of the struggling process (beyond the behavioral signals present in log data), focusing on why searchers struggled and where in the search session it became clear that their search task would succeed. We show that there are substantial differences in how searchers refine their queries in different stages in a struggling session. These differences have strong connections with session outcomes. There are particular pivotal queries (where it became clear the search would succeed) that play an important role in task completion. This pivotal query is often the last query and not all strategies are as likely to be pivotal. We developed classifiers to accurately predict key aspects of inter-query transitions for struggling searches, with a view to helping searchers struggle less.

We develop a classifier to predict query reformulation strategies during struggling search sessions. We show that we can accurately classify query reformulations according to an intent-based schema that can help select among different system actions. We also show that we can accurately identify pivotal queries within search sessions in which searchers are struggling.

Search engines aim to provide their users with the information that they seek with minimal effort. If a searcher is struggling to locate sought information, this can lead to inefficiencies and frustration. Better understanding these struggling sessions is important for designing search systems that help people find information more easily. We use our findings to propose ways in which systems can help searchers reduce struggling. Key components of such support are algorithms that accurately predict the nature of future actions and their anticipated impact on search outcomes. Our findings have implications for the design of search systems that help searchers struggle less and succeed more.

Acknowledgements. The first author performed this research during an internship at Microsoft Research and is supported by the Dutch national program COMMIT.

REFERENCES

[1] A. Hassan, R. W. White, S. T. Dumais, and Y.-M. Wang. Struggling or exploring? Disambiguating long search sessions. In *WSDM'14*, 2014.
[2] D. Odijk, R. W. White, A. Hassan Awadallah, and S. T. Dumais. Struggling and success in web search. In *CIKM 2015*, 2015.

*The full version of this paper appeared in CIKM2015 [2].

DIR 2015, November 27, 2015, Amsterdam, The Netherlands.

What Makes Book Search in Social Media Complex?

Marijn Koolen[1] Toine Bogers[3] Antal van den Bosch[4] Jaap Kamps[1,2]

[1] Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[2] Archives and Information Studies, University of Amsterdam, The Netherlands
{marijn.koolen,kamps}@uva.nl
[3] Department of Communication and Psychology, Aalborg University Copenhagen, Denmark
toine@hum.aau.dk
[4] Centre for Language Studies, Radboud University, The Netherlands
a.vandenbosch@let.ru.nl

This is a compressed abstract based on [2]. Real-world information needs are generally complex, yet almost all research focuses on either relatively simple search based on queries or recommendation based on profiles. It is difficult to gain insight into complex information needs from observational studies with existing systems; potentially complex needs are obscured by the systems' limitations. The general aim of this paper is to investigate whether explicit book search requests in social media can be used to gain insight in complex information needs that cannot be solved by a straightforward look-up search. We analyse a large set of annotated book requests from the LibraryThing discussion forums. We investigate 1) the comprehensiveness of book requests on the forums, 2) what relevance aspects are expressed in real-world book search requests, and 3) how different types of search topics are related to types of users, human recommendations, and results returned by retrieval and recommender systems.

We focus on LibraryThing[1] (LT), a popular social cataloguing site with 1.9 million members, which also offers a popular discussion forum to its users for discussing books. The forum is also used for book discovery: thousands of LT members use the forum to receive or provide recommendations for which books to read next. From the forums we can derive rich statements of requests, including explicit statements on the context of use and the context of the user, with example books and 'ground truth' human recommendations. We study this in the context of the INEX Social Book Search Track [1], which builds test collections around the book requests posted on the LT discussion forums. A non-random sample of 2,646 forum threads were annotated on whether the initial message is a *book request* or not, yielding 944 topics (36%) containing a book request. The requests were annotated with the relevance aspect(s) they express, based on set of 8 aspects: content-criteria (e.g. topic, genre), accessibility (reading level, length), familiarity (e.g. similar to previous reading experiences), novelty, engagement, metadata and socio-cultural resonance, with known-item searches labeled as a separate category. The books mentioned by other LT members in the thread were annotated as positive, neutral, or negative suggestions.

[1] `http://librarything.com/`, last accessed October 23, 2015.

DIR 2015, November 27, 2015, Amsterdam, Netherlands.

We found that the LT forum requests are comprehensive, with the majority containing multiple relevance aspects. The two dominating aspects are the *content* of the book (74% of requests) and looking for *familiar* reading experiences (36%), while other aspects are more oriented toward the reading context. In the majority of requests based on *familiarity* the requester provides example books to guide others in their recommendations. The *familiarity* aspects are also mostly used when searching fiction, whereas *content* aspects are less genre-specific. Finally, *content* aspects are more used by active users with large personal catalogues, where *familiarity* and contextual aspects are more typical of less active users with smaller catalogues. Suggestions for content and familiarity aspects also differ in terms of book popularity, with less popular books being suggested for content aspects than for familiarity aspects. The combination of content, context, and examples in a search request is a form of querying that is not supported by any current systems.

Retrieval systems can effectively use the content aspects of the search requests, and recommender systems can pick up signals in the requester's catalogue. With standard systems for retrieval (language model) and recommendation (K nearest neighbour) we demonstrated on a set of 680 requests and associated relevance judgements that a linear combination performs better than the individual approaches, in particular for topic groups where context and familiarity play a role. This suggest that the request type has an important role to play in the design of book discovery systems.

We highlighted the diversity of complex book search requests, and observed a mixture of content and context going beyond currently existing systems. Similar rich profiles and contextual information are available in many modern search scenarios. This is an important first step toward the development of novel information access systems that blend traditional search and (hybrid) recommendation approaches into a coherent whole.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (README project, NWO VIDI # 639.072.601; ExPoSe project, NWO CI # 314.99.108).

References

[1] M. Koolen, T. Bogers, J. Kamps, G. Kazai, and M. Preminger. Overview of the INEX 2014 social book search track. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, volume 1180 of *CEUR Workshop Proceedings*, pages 462–479. CEUR-WS.org, 2014.

[2] M. Koolen, T. Bogers, A. van den Bosch, and J. Kamps. Looking for books in social media: An analysis of complex search requests. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015. Proceedings*, pages 184–196, 2015.

Translation Model Adaptation Using Genre-Revealing Text Features (Abstract)

Marlies van der Wees Arianna Bisazza Christof Monz
Informatics Institute, University of Amsterdam
{m.e.vanderwees,a.bisazza,c.monz}@uva.nl

1. INTRODUCTION AND METHOD

We summarize the findings of van der Wees et al. [6]. Domain adaptation is an active field for statistical machine translation (SMT), and has resulted in various approaches that adapt system components to specific translation tasks. However, the concept of a *domain* is not precisely defined. Different domains typically correspond to different subcorpora, in which documents exhibit a particular combination of *genre* and *topic*. This definition has two major shortcomings: First, subcorpus-based domains depend on *provenance* information, which might not be available, or on manual grouping of documents into subcorpora, which is labor intensive and often carried out according to arbitrary criteria.

Second, the commonly used notion of a domain neglects the fact that topic and genre are two distinct properties of text [5]. While this distinction has long been acknowledged in text classification literature [3, 4, among others], most work on domain adaptation in SMT uses in-domain and out-of-domain data that differs on both the topic and the genre level, making it unclear whether the proposed solutions address topic or genre differences.

In our work, we follow text classification literature for definitions of the concepts topic and genre [4], and we recently studied the impact of both aspects on SMT [7]. Motivated by the observation that translation quality varies more between genres than across topics, we explore in this paper the task of *genre adaptation*. Concretely, we incorporate genre-revealing features, inspired by previous findings in genre classification literature, into a competitive translation model adaptation approach based on phrase pair weighting using a vector space model (VSM) [2].

In this approach, phrase pairs in the training data are represented by a vector capturing specific information about the phrase pair. In addition to the phrase pair vectors, a single vector is created for the development set which is similar to the test set. Next, for each training data phrase pair, we compute a similarity score between its vector and the development vector. This similarity is assumed to indicate the relevance of the phrase pair with respect to the test set's genre and is added to the decoder as a new feature.

We compare a number of variants of the general VSM framework, differing in the way vectors are defined and constructed. First, we adhere to the common scenario in which adaptation is guided by manual subcorpus labels that resemble the training data's provenance. Next, to move away from manual labels, we explore the use of genre-revealing features that have proven successful for distinguishing genres in classification tasks. The features that are most discriminative between the genres in our test sets (newswire (NW) and user-generated (UG) text) are counts of first and second person pronouns, exclamation and question marks, repeating punctuation, emoticons, and numbers. Finally, we use LDA-inferred [1] document distributions as a third vector representation.

2. RESULTS AND CONCLUSIONS

We evaluate different VSM variants on two Arabic-to-English translation tasks, both comprising the genres NW and UG. Besides VSM variants containing only one of the presented feature types (i.e., manual provenance labels, automatic genre features, or LDA distributions), we also explore various combinations in which multiple VSM similarities are added as additional decoder features. Our best performing system includes both genre features and LDA distributions, suggesting that the two vector representations are to some extent complementary.

In a series of experiments we show that automatic indicators of genre can replace manual subcorpus labels, yielding significant improvements of up to 0.9 BLEU over a competitive unadapted baseline. In addition, we observe small improvements when using automatic genre features on top of manual subcorpus labels. We also find that the genre-revealing feature values can be computed on either side of the training bitext, indicating that our proposed features can be robustly projected across languages. Therefore, the advantages of using the proposed method are twofold: (i) manual subcorpus labels are not required, and (ii) the same set of features can be used successfully across different test sets and languages. Finally, we find that our genre-adapted translation models encourage document-level translation consistency (i.e., consistent translation of repeated phrases within a single document) with respect to the unadapted baseline.

REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
[2] B. Chen, R. Kuhn, and G. Foster. Vector space model for adaptation in statistical machine translation. In *Proceedings of ACL*, pages 1285–1293, 2013.
[3] N. Dewdney, C. VanEss-Dykema, and R. MacMillan. The form is the substance: classification of genres in text. In *Proceedings of the Workshop on HLT and KM*, 2001.
[4] M. Santini. State-of-the-art on automatic genre identification. Techn. Report ITRI-04-03, Information Technology Research Institute, University of Brighton, 2004.
[5] B. Stein and S. Meyer Zu Eissen. Distinguishing topic from genre. In *Proceedings of I-KNOW 06*, pages 449–456, 2006.
[6] M. van der Wees, A. Bisazza, and C. Monz. Translation model adaptation using genre-revealing text features. In *Proceedings of DiscoMT'15*, pages 132–141, 2015.
[7] M. van der Wees, A. Bisazza, W. Weerkamp, and C. Monz. What's in a domain? Analyzing genre and topic differences in statistical machine translation. In *Proceedings of ACL*, pages 560–566, 2015.

Image2Emoji: Zero-shot Emoji Prediction for Visual Media

Abstract for Dutch Belgium Information Retrieval Workshop 2015

Spencer Cappallo[†] Thomas Mensink[†] Cees G. M. Snoek[†‡]

[†]University of Amsterdam [‡]Qualcomm Research Netherlands

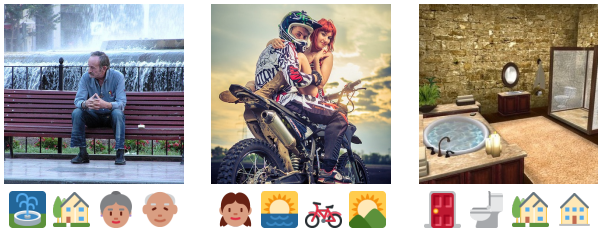

Figure 1: Emoji predictions for three images.

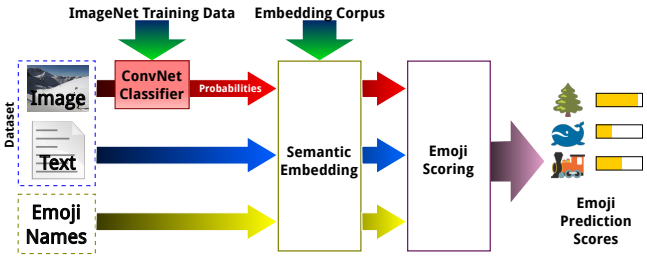

Figure 2: Image2Emoji data flow. Probability scores from visual classifiers along with the image's accompanying text are placed in the semantic embedding. Their similarity to the target emoji label names are used to score the emoji.

We present *Image2Emoji* [1], a multi-modal approach for generating ideogram representations of images in a zero-shot manner (*i.e.*, without labeled examples of imagery with emoji). Ideogram-based representations have several advantages over textual or natural image representations. As visual iconography, they maintain a visual grammar of interaction for queries on visual media. Their nature is inherently language independent and accessible by the full spectrum of age groups. These representations can be subsequently applied to a variety of search and summarization tasks.

Emoji are used as our target ideogram set. Emoji are a set of hundreds of ideograms representing a wide spread of semantic concepts, which have been adopted and implemented into all major smart phone platforms, and most major social media websites. As a unicode-specified set of ideograms, their extent and mapping is platform agnostic. Furthermore, emoji are increasingly used for communication on social media, which means that the userbase arrives with a pre-installed familiarity to the available ideograms. The spread of emoji as a means of communication also suggests that they are semantically diverse and rich enough to facilitate description of a wide range of scenarios and concepts. Lastly, the small-but-legible design and square form factor of emoji make them an interesting candidate for interfaces on small screens, such as smart watches.

We employ a multi-modal, zero-shot approach to generating emoji description, which relies on a shared semantic embedding space in which all modalities and the target modality (emoji) can be related. Distinct from related zero-shot image-to-text approaches, we exploit both visual and textual features to make our zero-shot predictions. These multi-modal features are embedded in a word2vec [4] space trained on an accompanying text of 100M photos from Flickr [5], and their proximity to the embedding of the emoji within this space is used to calculate the final emoji representation. Our multi-modal approach is experimentally validated in a more traditional setting through zero-shot classification on the MSCOCO [3] train dataset, which consists of 80,000 images annotated with 80 classes.

We present subjective results for three potential applications: query-by-emoji, emoji ranking, and emoji summarization. Query-by-emoji utilizes emoji as building blocks for composing searches for visual data, emoji ranking allows for short emoji-based representations of visual media, and emoji summarization uses representations to summarize a collection of related visual documents, such as an image album or a video. In a closely related technical demo [2], we provide an example of a video search engine using query-by-emoji and returning emoji summaries of the video contents.

1. REFERENCES

[1] S. Cappallo, T. Mensink, and C. G. M. Snoek. Image2emoji: Zero-shot emoji prediction for visual media. In *MM*, 2015.

[2] S. Cappallo, T. Mensink, and C. G. M. Snoek. Query-by-emoji video search. In *MM*, 2015.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[5] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv:1503.01817*, 2015.

Dutch Belgium Information Retrieval Workshop 2015
Original publication [1] published at ACM MM'15

Learning to Combine Sources of Evidence for Indexing Political Texts

Extended Abstract

Mostafa Dehghani[1] Hosein Azarbonyad[2] Maarten Marx[2] Jaap Kamps[1]

[1]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[2]Informatics Institute, University of Amsterdam, The Netherlands
{dehghani,h.azarbonyad,kamps,maartenmarx}@uva.nl

ABSTRACT

Political texts are pervasive on the Web and access to this data is crucial for government transparency and accountability to the population. However, access to such texts is notoriously hard due to the ever increasing volume, complexity of the content and intricate relations between these documents. Indexing documents with a controlled vocabulary is a proven approach to facilitate access to these special data sources. However, increasing production of the political text makes human indexing very costly and error-prone. Thus, technology-assisted indexing is needed which scale and can automatically index any volume of texts.

There are different sources of evidence for the selection of appropriate indexing terms for political documents, including variant document and descriptor term representations, the structure of thesauri, if existing, and the set of annotated documents with the descriptor terms assigned as training data.

The main goal of this research is to investigate the effectiveness of different sources of evidence—such as the labeled training data, textual glosses of descriptor terms, and the thesaurus structure—for indexing political texts and combine these sources to have a better performance. We break down our main goal into three concrete research questions:

RQ1 *How effective is a learning to rank approach integrating a variety of sources of information as features?*

RQ2 *What is the relative importance of each of these sources of information for indexing political text?*

RQ3 *Can we select a small number of features that approximate the effectiveness of the large LTR system?*

We make use of learning to rank (LTR) as a means to not only take advantage of all sources of evidence effectively, but also analyse the importance of each source. To do so, we consider each document to be annotated as a query, and using all text associated with a descriptor term as documents. We evaluate the performance of the proposed LTR approach on the English version of JRC-Acquis [3] and compare our results with JEX [4] which is one of the state of the art systems developed for annotating political text.

1. EXPERIMENTS AND ANALYSIS

For our experiments, we have used English documents of last five years (from 2002 to 2006) of JRC-Acquis dataset [3]. The documents of this corpus have been manually labeled with EuroVoc concepts [2].

To address RQ1, using a LTR approach for integrating all features, we observe significantly better performance than previous systems. Table 1 shows the evaluation results of the proposed method compared to the baseline systems. Furthermore, we define features based on similarity of titles, texts, and descriptions between documents and descriptor terms as well as structural features, i.e. popularity, ambiguity, and generality, and then use LTR as a analytic tool to address our second research question. The analysis of feature weights reveals the relative importance of various sources of evidence, also gives insight in the underlying classification problem (Figure 1). Finally, based on the analysis of feature importance, we study RQ3. We suggest a lean-and-mean system using only four features (text, title, descriptor glosses, descriptor term popularity) which is able to perform at 97% of the large LTR model. The result of the lean-and-mean system is also presented in Table 1.

Table 1: Performance of JEX, best single feature, LTR, and lean-and-mean system. We report "incremental" improvement and significance ($^\blacktriangle$ indicates t-test, one-tailed, p-value < 0.05)

Method	P@5 (%Diff.)	Recall@5 (%Diff.)
JEX	0.4353	0.4863
BM25-TITLES	0.4798 (10%)$^\blacktriangle$	0.5064 (4%)$^\blacktriangle$
LTR-ALL	0.5206 (9%)$^\blacktriangle$	0.5467 (8%)$^\blacktriangle$
LTR-TTGP	0.5058 (-3%)	0.5301 (-3%)

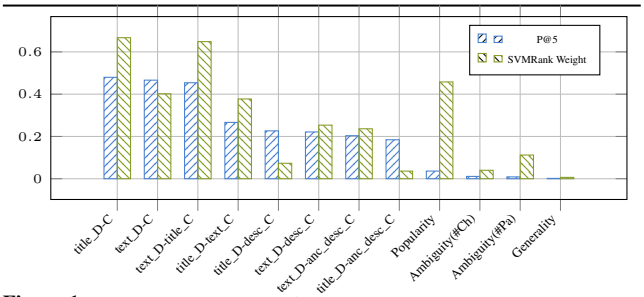

Figure 1: Feature importance: 1) $P@5$ of individual features, 2) weights in SVMRank model

Acknowledgments The full version of this paper is available as [1]. This research is funded by the Netherlands Organization for Scientific Research (WebART project, NWO CATCH # 640.005.001; ExPoSe project, NWO CI # 314.99.108; DiLiPaD project, NWO Digging into Data # 600.006.014).

References

[1] M. Dehghani, H. Azarbonyad, M. Marx, and J. Kamps. Sources of evidence for automatic indexing of political texts. In *Proceedings ECIR*, pages 568–573, 2015. URL http://dx.doi.org/10.1007/978-3-319-16354-3_63.

[2] EuroVoc. Multilingual thesaurus of the european union. http://eurovoc.europa.eu/.

[3] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufis. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *LREC*, pages 2142–2147, 2006.

[4] R. Steinberger, M. Ebrahim, and M. Turchi. JRC EuroVoc indexer JEX-A freely available multi-label categorisation tool. In *LREC*, pages 798–805, 2012.

Automatically Assessing Wikipedia Article Quality by Exploiting Article–Editor Networks*

Xinyi Li, Maarten de Rijke
{x.li,derijke@uva.nl}

University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

We consider the problem of automatically assessing Wikipedia article quality. We develop several models to rank articles by using the editing relations between articles and editors. First, we create a basic model by modeling the article-editor network. Then we design measures of an editor's contribution and build weighted models that improve the ranking performance. Finally, we use a combination of featured article information and the weighted models to obtain the best performance. We find that using manual evaluation to assist automatic evaluation is a viable solution for the article quality assessment task on Wikipedia.

1. INTRODUCTION

Wikipedia is the largest online encyclopedia built by crowdsourcing, on which every-one is able to create and edit the contents. Its articles vary in quality and only a minority of them are manually evaluated high quality articles. Since manually labeling articles is inefficient, it is essential to automatically assess article quality. Our task is motivated by the assumption that automatic procedures for assessing Wikipedia article quality can help information retrieval that utilizes Wikipedia resources and information extraction on Wikipedia to obtain high quality information.

We develop several models to rank articles by quality. Our first motivation is to see if the importance of a node in the network can indicate quality. So we develop a basic PageRank-based model. Additionally, instead of treating links as equal in the basic model, we tweak the model by putting weights on the links to reflect the difference of editor contributions. Finally, we utilize existing manual evaluation results to improve automatic evaluation.

2. EXPERIMENTS

We assess article quality by ranking. Since featured articles are the best quality articles on Wikipedia, they are frequently used as the gold standard to measure ranking performance. We consider recall scores at the first N items in the result set, as well as precision-recall curves.

We address two main research questions. We contrast our four methods, i.e., the Baseline method, the simple weighted model (SW), the complex weighted (CW) model, as well as two variants with probabilistic initial values (SWP, CWP) in Figure 1. To determine whether the observed differences between two models are statistically significant, we use Student's t-test, and look for significant improvements (two-tailed) at a significance level of 0.99. We find that both SWP and CWP statistically significantly outperform other models in all categories. We also note that the SWP model performs better than the CWP model in most cases, which is contrary to the previous experiments where the complex contribution measure yields better results. The best ranking performance is achieved by the SWP model when using all available high quality articles in initialization. And the recall levels are up to an applicable value. E.g., the recall value at $N = 200$ is 0.756 in geography, meaning that the 180 featured articles in that category have a probability of 75.6% to appear in the top-200 list.

*An extended version of this abstract paper has been accepted at ECIR 2015 Li et al. [1].

3. REFERENCES

[1] X. Li, J. Tang, T. Wang, Z. Luo, and M. de Rijke. Automatically assessing wikipedia article quality by exploiting article–editor networks. In *ECIR 2015: 37th European Conference on Information Retrieval*, pages 574–580. Springer, 2015.

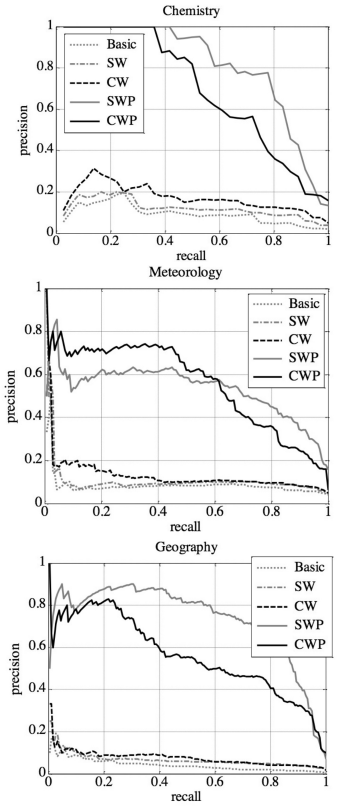

Figure 1: Precision-recall curves for the baseline (Basic), simple weighted (SW), complex weighted (CW), simple weighted probabilistic (SWP), complex weighted probabilistic (CWP) model.

A Hybrid Approach to Domain-Specific Entity Linking

[Extended Abstract]

Alex Olieman Jaap Kamps Maarten Marx Arjan Nusselder
University of Amsterdam, Amsterdam, The Netherlands
{olieman|kamps|maartenmarx}@uva.nl arjan@nusselder.eu

1. INTRODUCTION

In the Entity Linking (EL) task, textual mentions are linked to corresponding Knowledge Base (KB) entries. The majority of state-of-the-art EL systems utilize one or more open-domain KBs, such as Wikipedia, DBpedia, Freebase, or YAGO, as basis for learning their entity recognition and disambiguation models [3]. The results of annotating a domain-specific corpus disappoint, however, when using a domain-agnostic EL system. We propose to use specialized linkers for salient entity types within the corpus' domain, which can work in concert with a generally trained model. Our approach is applied to conversational text, in particular parliamentary proceedings. The techniques that we have investigated are designed to be applicable to written records of any kind of conversation.

2. DOMAIN-SPECIFIC ENTITY LINKING

The specialist linkers are developed to target specific entity types that are mentioned frequently in the target corpus. These linkers capitalize on a small amount of background knowledge, and achieve entity recognition and disambiguation by means of pattern detection, string matching, and structured queries against the corpus. The simplest way that we have considered to annotate entities of a specific type is based on exact string matching. In our corpus we target Dutch political parties (*n=155*), because they are highly relevant as well as unambiguously named.

Our second linker applies to ambiguous entity types, and targets mentioned persons. It utilizes information about which people were present during a conversation, and about the period(s) during which a person was active, for disambiguation. Government and parliament members (*n=3,664*) are targeted in our corpus, and some knowledge of debating etiquette assists in detecting where they are mentioned. We also detect where government members are mentioned by their role, by means of a temporal index which maps roles to persons.

3. BENCHMARK

We have selected a sample of Dutch parliamentary proceedings from the period 1999–2012, which was subsequently annotated by DBpedia Spotlight (DBpS) [1], UvA Semanticizer (F+S) [2], and the specialist linkers. In order to assess the quality of these annotations against a consistent gold standard, we employed two human annotators for an independent and a consensus-building annotation round. To combine the output of multiple systems, we employ a preference ordering: the most specialized (i.e. estimated high-precision) system is asked to link a phrase first, and only if it doesn't the second system in the order is asked, and so on. By adding a generalist EL system at the end of the chain, the phrases that mention non-domain-specific entities also have their chance at being linked.

4. RESULTS

The results show that the specialist linkers were able to generate a larger number of accurate annotations for the corpus than either of the baseline systems, whilst limited to two specific entity types. F+S is the more precise of the baselines, but DBpS produces a greater number of potentially useful links. Our approach of combining a relatively simple custom-made EL system with an off-the-shelf EL system has also proven to be successful. This combination strategy produced a significantly better result than any of the systems could by themselves.

5. CONCLUSIONS

The current state-of-the-art entity linking systems aim to be open-domain solutions for corpora that are as heterogeneous as the Web. An unfortunate effect of this aim is that such generalist EL systems often disappoint when they are used on domain-specific corpora. We have outlined the prerequisites for, and development of, a lightweight linking system that targets salient entity types in a specific corpus. The specialist system, two baseline generalist systems, and hybrid combinations thereof have been evaluated against a gold standard, which is available as an open-data benchmark for the EL community at `http://datahub.io/dataset/el-bm-nl-9912`.

Our results show that the specialist system offers competitive performance to the two baseline systems, even though it is limited to two highly specific entity types. Moreover, by combining the specialist linkers with one or both generalist EL systems, recall can be significantly increased at a modest precision cost.

6. REFERENCES

[1] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proc. of I-Semantics 2013*, pages 3–6, Austria, Graz, 2013.

[2] D. Odijk, E. Meij, and M. de Rijke. Feeding the Second Screen: Semantic Linking based on Subtitles. In *OAIR 2013*, 2013.

[3] W. Shen, J. Wang, and J. Han. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 4347(2):443–460, 2014.

Using Logged User Interactions for Ranker Evaluation

Artem Grotov
a.grotov@uva.nl

Shimon Whiteson
s.a.whiteson@uva.nl

Maarten de Rijke
derijke@uva.nl

University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

We address the problem of how to safely compare rankers for information retrieval. In particular, we consider how to control the risks associated with switching from an existing *production ranker* to a new *candidate ranker*. Whereas existing online comparison methods require showing potentially suboptimal result lists to users during the comparison process, which can lead to user frustration and abandonment, our approach only requires user interaction data generated through the natural use of the production ranker. Specifically, we propose a Bayesian approach for (1) comparing the production ranker to candidate rankers and (2) estimating the confidence of this comparison. The comparison of rankers is performed using click model-based information retrieval metrics, while the confidence of the comparison is derived from Bayesian estimates of uncertainty in the underlying click model [1]. These confidence estimates are then used to determine whether a risk-averse decision criterion for switching to the candidate ranker has been satisfied. Experimental results on several learning to rank datasets and on a click log show that the proposed approach outperforms an existing ranker comparison method that does not take uncertainty into account [2].

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Ranker evaluation; Learning to rank; Click models

Acknowledgements. This research was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, CI-14-25, SH-322-15, Amsterdam Data Science, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

REFERENCES

[1] A. Grotov, A. Chuklin, I. Markov, L. Stout, F. Xumara, and M. de Rijke. A comparative study of click models for web search. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 78–90. CLEF, 2015.

[2] A. Grotov, S. Whiteson, and M. de Rijke. Bayesian ranker comparison based on historical user interactions. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 273–282, New York, NY, USA, 2015. ACM.

KISS MIR: Keep It Semantic and Social Music Information Retrieval

Amna Dridi
Faculty of Computer Science
Free University of Bozen-Bolzano
Piazza Domenicani 3, Bozen-Bolzano (39100)
amna.dridi@inf.unibz.it

ABSTRACT

While content-based approaches for music information retrieval (MIR) have been heavily investigated, user-centric approaches are still in their early stage. Existing user-centric approaches use either music-context or user-context to personalize the search. However, none of them give the possibility to the user to choose the suitable context for his needs. In this paper we propose KISS MIR, a versatile approach for music information retrieval. It consists in combining both music-context and user-context to rank search results. The core contribution of this work is the investigation of different types of contexts derived from social networks. We distinguish semantic and social information and use them to build semantic and social profiles for music and users. The different contexts and profiles can be combined and personalized by the user. We have assessed the quality of our model using a real dataset from Last.fm. The results show that the use of user-context to rank search results is two times better than the use of music-context. More importantly, the combination of semantic and social information is crucial for satisfying user needs.

General Terms

IR Theory and Practice

Keywords

music information retrieval, music-context, user-context, personalization, social information

1. REFERENCES

[1] T. Arjannikov, C. Sanden, J.Z. Zhang. "Verifying Tag Annotations through Association Analysis," *Proceedings of ISMIR*, pp. 195–200, 2013.

[2] D. Boland, R. Murray-Smith. "Information-Theoretic Measures of Music Listening Behaviour," *Proceedings of ISMIR*, pp. 561–566, 2014.

[3] D. Bountouridis, R.C. Veltkamp, Y.V. Balen. "Placing Music Artists and Songs in Time using Editorial Metadata and Web Mining Techniques," *ISMIR*, 2013.

[4] D. Bugaychenko, A. Dzuba. "Musical Recommendations and Personalization in a Social Network," *Proceedings of RecSys'13*, pp. 367–370, 2013.

[5] Z. Chedrawy, S. Abidi. "A Web Recommender System for Recommending Predicting and Personalizing Music Palylists," *Proc. of WISE*, pp. 335–342, 2009.

[6] H.C. Chen, A. L. P. Chen. "A Music Recommendation System Based on Music Data Grouping and User Interests," *Proceedings of CIKM*, pp. 231–238, 2001.

[7] J.S. Downie. "The Scientific Evaluation of Music Information Retrieval Systems: Foundation and Future," *Computer Music Journal*, pp. 12–23, 2004.

[8] P. Knees, T. Pohle, M. Schedl, G. Widmer. "A Music Search Engine Built upon Audio-based and Web-based similarity measures," *SIGIR*, pp. 447–454, 2007.

[9] P. Knees, M. Schedl, O. Celma. "Hybrid Music Information Retrieval," *Proceedings of ISMIR*, pp. 1–2, 2013.

[10] T. Li, M. Ogihara. "Toward Intelligent Music Retrieval," *Proceedings of MULTIMEDIA*, Vol. 8, No. 3, pp. 564–574, 2006.

[11] M. Schedl, D. Hauger. "Mining microblogs to infer music artist similarity via web mining techniques," *Proceedings of WWW*, pp. 877–886, 2012.

[12] Y.Song, S. Dixon, M. Pearce, A. Halpern. "Do Online Social Tags predict perceived or induced emotional responses to Music?," *Proc. of ISMIR*, pp. 89–94, 2013.

[13] B. Zhang, J. Shen, Q. Xiang, Y. Wang. "CompositeMap: a Novel Framework for Music Similarity," *Proceedings of SIGIR*, pp. 403–410, 2009.

A Search Engine with Reading Level Specific Results

Thijs Westerveld
WizeNoze
thijs@wizenoze.com

ABSTRACT
This paper provides a description of jouwzoekmachine.nl, a search engine designed for children. The engine differentiates search results by reading level and helps children formulate their queries.

Keywords
Web Search, Search Engine for Children, Readability Classification, JouwZoekmachine.nl

1. INTRODUCTION
Children use the internet from a very young age. 78 percent of Dutch toddlers and pre-schoolers and 5 percent of babies younger than one year are going online [2]. Unfortunately, the online content nor the search tools to which children have access have been developed with children in mind. As a result, children are often exposed to content that is inappropriate in many ways. It may be harmful or too complex. JouwZoekmachine.nl changes this[1]. It is the first search engine for children that takes the reading level of the user into account. Children can use the engine in school and at home for informational search at their own level. This paper outlines the main technology supporting this search engine.

2. SOURCE SELECTION
The documents in jouwzoekmachine.nl come from carefully selected sources. The manual curration of the sources is important to make sure the content we disclose is safe, trustworthy, and aimed at children between 6 and 15 years old. We explicitly exclude shopping pages, sources with many advertisements, information aimed at parents or teachers, and information written *by* children.

Many sources treat children as one homogeneous group, but there are huge differences between the reading level of for example a 7 years old child, who is only just learning to read, and a 13 year old child, who finished primary education. To facilitate reading level specific search we automatically determine the level of each document that we collect.

3. TEXT CLASSIFICATION
The automatic identification of the reading level of a text has been studied since the 1940. Early metrics are based on linear combinations of shallow text features like average word and sentence length, average number of syllables and fraction of common term. More recently, readability classification has been treated as a language modeling or machine learning problem. For an overview of readability research, see [1]. We take the machine learning route, extracting shallow text features as well as lexical and vocabulary features and building a readability classifier from labeled training data. A wide variety of training data is used, including text books, web pages, and news for children. To train a classification model for this heterogeneous dataset, we have to deal with a variety of labels and class granularities. We map all labels to an internal difficulty scale to be able to train a single classification model across all source types.

4. LABEL AGREEMENT
To test the reliability of the labels in our training data, we set up an experiment to assess the agreement between multiple assessors on the readability of a text[2]. In this experiment we ask participants to assess the reading level of 10 short text fragments taken from the BasiLex corpus [3]. They have to indicate at what age they think an average child can read the presented fragment. Preliminary results based on the first 63 participants (333 documents with at least two assessments) show that the agreement is very low. In only 23.7% of the cases assessors agree on the exact reading level of a fragment, and for only 50.8% of the fragments the readability assessments are at most one year apart.

5. FEEDBACK FROM CHILDREN
To further improve our classifiers, we plan to retrain them on data labeled by children. We will collect feedback from children both explicitly through a readability game and implicitly through their interaction with jouwzoekmachine.nl. With these data we expect to improve our classification and ranking algorithms. This way, we make sure the internet does not only get safer for children, but also more suitable and more readable.

6. REFERENCES
[1] K. Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2), 2014.

[2] R. Pijpers. App noot muis: Peuters en kleuters op internet, 2011.

[3] A. Tellings, M. Hulsbosch, A. Vermeer, and A. van den Bosch. Basilex: an 11.5 million words corpus of dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, 4:191–208, 12/2014 2014.

[1] http://jouwzoekmachine.nl

[2] http://leesbaarheidstest.wizenoze.com

QUINN: Query Updates for News Monitoring

Suzan Verberne
Radboud University, The Netherlands

Thymen Wabeke
TNO, The Hague, The Netherlands

Rianne Kaptein
TNO, The Hague, The Netherlands

LexisNexis Publisher[1] is an online tool for news monitoring. Organizations use the tool to collect news articles relevant to their work. For monitoring the news for a user-defined topic, LexisNexis Publisher takes a Boolean query as input, together with a news collection and a date range. The output is a set of documents from the collection that match the query and the date range.

For the users it is important that no relevant news stories are missed. Therefore, the query needs to be adapted when there are changes to the topic. This can happen when new terminology becomes relevant for the topic (e.g. 'wolf' for the topic 'biodiversity'), there is a new stakeholder (e.g. the name of the new minister of economic affairs for the topic 'industry and ICT') or new geographical names are relevant to the topic. (e.g. 'Heumensoord' for the topic 'refugees') The goal of the current work is to support users of news monitoring applications by providing them with suggestions for query modifications in order to retrieve more relevant news articles.

The user can control the precision of the final publication list by disregarding irrelevant documents in the selection. Recall is more difficult to control because the user does not know what he has not found. Our intuition is that documents that are relevant but *not* retrieved with the current query have similarities with the documents that *are* retrieved by the current query. Therefore, our approach to query suggestion is to generate candidate query terms from the set of retrieved documents. This approach is related to pseudo-relevance feedback, a method for query expansion that assumes that the top-k retrieved documents are relevant, extracting terms from those documents and adding them to the query. There are three key differences with our approach: First, instead of adding terms blindly, we provide the user with suggestions for query adaptation. Second, we take into account an important characteristic of news data: the collection is constantly changing. We hypothesize that terms that show a big increase in frequency over time are candidate new query terms, because they were not relevant in an earlier stage of the news stream. Third, we have to deal with Boolean queries, which implies that we do not have a relevance ranking of documents to extract terms from. This means that the premise of 'pseudo-relevance' may be weak for the set of retrieved documents.

Our approach for query term extraction is as follows: For a given Boolean query, we retrieve the result set R_{recent}, which is the set of articles published in the last 30 days, and the result set R_{older}, which is the set of articles published 60 to 30 days ago. We implemented four different term scoring algorithms from the literature, and used each of them to extract three term lists: T_1 is the divergence between R_{recent} and a generic news background corpus; T_2 is the divergence between R_{recent} and R_{older}; T_3 is the divergence between R_{older} and the generic news background corpus. The query suggester returns one of three term lists to the user: $A = T_1$; $B = T_2$ and $C = \{t : t \in T_1 \wedge t \notin T_3\}$.

The demo application has been used to collect feedback from expert users of LexisNexis Publisher to determine the best method for generating term suggestions. In the application, a Boolean query can be entered that is used to search in Dutch newspapers. The found documents are shown in a result list and a list of query term suggestions (a pool of terms from all methods) is presented. Users were asked to judge the relevance of the returned terms on a 5-point scale, could update the search query (potentially with a suggested term) and retrieve a new result list.

The results of our user experiment show that with the best performing method (method A with either Parsimonious Language Models or Kullback-Leibler Divergence as term scoring algorithm), the user selected a term from the top-5 suggestion list for only 13% of the topics, and judged at least one term as relevant (relevance score >= 4) for 25% of the topics. Inspection of the results and the user comments revealed that the term suggestions are noisy, mainly because the set of retrieved documents for the Boolean query is noisy. We expect that the use of relevance ranking instead of Boolean retrieval, and a post-filtering for noisy terms, will give better user satisfaction.

Acknowledgements

This publication was supported by the Dutch national program COMMIT (project P7 SWELL).

[1]http://www.lexisnexis.com/bis-user-information/publisher/

BioMed Xplorer: A tool for exploring (bio)medical knowledge

Mohammad Shafahi
Informatics Institute
Sciencepark 904
Amsterdam, The Netherlands
m.shafahi@uva.nl

Hamideh Afsarmanesh
Informatics Institute
Sciencepark 904
Amsterdam, The Netherlands
h.afsarmaensh@uva.nl

Hayo Bart
Informatics Institute
Sciencepark 904
Amsterdam, The Netherlands
hayojay@live.nl

1. BIOMED XPLORE

Developing an effective patients' risk prediction model for a disease, as commonly needed by practitioners for assessing patients, requires exploration of a vast body of published (bio)medical knowledge. Furthermore, exponential growth of this body of knowledge, as indicated in [2] for the MEDLINE biomedical bibliographic database, poses challenges to this knowledge exploration effort. As of today, this database contains over 22 million citations, where over 750,000 of which were added in 2014 [5].Numerous researchers have attempted to address this issue by developing different approaches and support tools [4, 1], for example through developing comprehensive visualization of the knowledge extracted from (bio)medical publications [6, 7, 3]. Most of these however, do not sufficiently support the needed dynamism in the body of this knowledge, lack intuitiveness in their use, and present a rather small amount of information which is usually obtained from a single source, whereas further information, related to the same concept, can be obtained from multiple external sources.

BioMed Xplorer aims to address these gaps through the use of a dynamic model of (bio)medical knowledge, represented as a network of interrelated (bio)medical concepts, and integrating disperse sources across the web[8]. Additionally, semantic web technologies are incorporated into the tool, to better support handling of large amounts of available dynamic and heterogeneous information. BioMed Xplorer is an interactive tool enabling biomedical researchers to explore the needed body of knowledge and its provenance data, by modeling the biomedical concepts and their relationships in an information graph.

Using BioMed Xplorer, researchers can explore knowledge about a disease through a user friendly and intuitive interface. Furthermore, it provides disease related information through a multitude of sources, while preserving and presenting their provenance data. The knowledge currently represented by the BioMed Xplorer primarily reflects knowledge from SemMedDB, which is a large relational SQL database, conceptualizing relationships among (bio)medical concepts, and extracted from the PubMed articles. In BioMed Xplorer, an RDF knowledge base is created that maps SemMedDB concepts. The mapping is conducted based on a core ontology introduced in our approach.

2. REFERENCES

[1] Aaron M Cohen and William R Hersh. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71, 2005.

[2] Lawrence Hunter and K Bretonnel Cohen. Biomedical language processing: what's beyond pubmed? *Molecular cell*, 21(5):589–594, 2006.

[3] Halil Kilicoglu, Marcelo Fiszman, Alejandro Rodriguez, Dongwook Shin, A Ripple, and Thomas C Rindflesch. Semantic medline: a web application for managing the results of pubmed searches. In *Proceedings of the third international symposium for semantic mining in biomedicine*, pages 69–76, 2008.

[4] Zhiyong Lu. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 2011.

[5] U.S. National Library of Medicine. Medline fact sheet. `http://www.nlm.nih.gov/pubs/factsheets/medline.html`. [Online; accessed August 28 2015].

[6] Conrad Plake, Torsten Schiemann, Marcus Pankalla, Jörg Hakenberg, and Ulf Leser. Alibaba: Pubmed as a graph. *Bioinformatics*, 22(19):2444–2445, 2006.

[7] Dietrich Rebholz-Schuhmann, Harald Kirsch, Miguel Arregui, Sylvain Gaudan, Mark Riethoven, and Peter Stoehr. Ebimed-text crunching to gather facts for proteins from medline. *Bioinformatics*, 23(2):e237–e244, 2007.

[8] Mohammad Shafahi, Hayo Bart, and Hamideh Afsarmanesh. Biomed xplorer: Exploring (bio)medical knowledge using linked data. In *Proceedings of the Seventh International Conference on Bioinformatics Models, Methods and Algorithms (accepted)*, 2016.

Let the Children Play - Developing a Child Feedback Collection System for Text Readability Assessments

Rutger Varkevisser*
WizeNoze
rutger@wizenoze.com

Theo Huibers
University of Twente
t.w.c.huibers@utwente.nl

Thijs Westerveld
WizeNoze
thijs@wizenoze.com

ABSTRACT

This paper describes a project which has the goal of improving the classification of textual content for children via a crowd-sourced gamification method. In this demonstration we present early prototypes which focus on different methods of interaction.

1. INTRODUCTION

Children these days are very connected to the internet and have the ability to access a wealth of information. However, there is not much age-specific relevant information available (besides games and videos), and when there is, it is buried deep within the results of most search-engines which makes it almost impossible to find.

This creates an urgent need to inform children with age-specific information. Since texts which are too complex are too difficult for a child to understand, and texts which are too simple are seen as boring[1]. This then poses the following question: *how do we determine the age-appropriateness of these text fragments?* Classification of text fragments can be done via several methods e.g. automatically using a classification system, by the content owner itself or by a specialized group to name a few.

Even with these methods at our disposal, it remains very difficult to effectively determine the age-appropriateness of text fragments and therefore classify them correctly. To assist the current systems WizeNoze[1] uses in the classification of texts, the current project was set-up.

2. AIM OF THE PROJECT

This project attempts to improve the classification of textual content by letting children play with the text in such a way that allows the system to determine the appropriateness and readability of the text for the user, whom the system knows has a specific age- and education level. By presenting the same text to more and more children and combining readability scores for all of them, the system is able to gather large amount of data about the classification of individual text fragments. This data can then be used as training data for the WizeNoze classification systems. This all contributes to the ultimate goal of providing a more accurate estimation of the readability level of texts.

*Master student at the University of Twente, currently doing an internship at WizeNoze

[1] `www.wizenoze.com`

3. PROTOTYPES AND USER-TEST

The project was started in begin September 2015, and recently the first user-test was completed using several prototypes. These prototypes were designed to use (digital) variations of the Cloze-test procedure[2] where, at a certain interval within a text, words are omitted and replaced by a blank line for the user to fill in. The number of correct answers given can be converted to a percentage based score which determines the readability of the text for the users age/education level.

For the first user-test we developed three prototypes with each its own unique method of interaction. The first prototype uses a drag-and-drop system, the second a multiple choice system and the third uses exact matching method to select/fill-in the correct answers. The first user-test was a small scale qualitative test to determine the suitability of the various prototypes and test the different interaction methods, with the results providing feedback that will be used to improve the current system and prepare for the next phase of user-testing. The main focus of these future user-tests will be to determine whether the interactive (digital) versions of the Cloze-test can be made fun without losing reliability in the readability scores they provide.

Results from the first user-test showed that the exact matching version was considered as too hard and that the participating children preferred the drag-and-drop and multiple choice versions. While qualitative in nature and without claiming significance, data gathered from the first user-test seem to suggest a loose correlation between the number of correct results and completion time between each prototype.

4. DEMONSTRATION

In the demonstration the latest version of the prototypes are shown and the approach of the project is further explained. Additionally the future heading of this project is discussed. Online versions of the above mentioned prototypes are made available at `dir2015.rutgerv.com` for the attendees to play/interact with.

5. REFERENCES

[1] F. Sluis, E. L. Broek, R. J. Glassey, E. M. Dijk, and F. M. Jong. When complexity becomes interesting. *Journal of the Association for Information Science and Technology*, 65(7):1478–1500, 2014.

[2] W. L. Taylor. Cloze procedure: a new tool for measuring readability. *Journalism quarterly*, 1953.

Multilingual Word Embeddings from Sentence Representations

Benno Kruit
benno.kruit@student.uva.nl

Sara Veldhoen
sara.veldhoen@student.uva.nl

Motivation.

Distributional semantics concerns the establishment of a semantic space where words are represented by vectors and their relations have a geometrical interpretation. We investigated how to use data from multiple languages to create a single semantic space. In particular, we induce word embeddings for seven languages in a shared space.

This allows for linguistic inquiry into semantic differences between vocabularies. The latent conceptual semantic space can arguably be approximated more closely by using multilingual data, as language-specific effects fade away. Moreover, a joint semantics space of many languages allows for tasks such as multilingual Information Retrieval without using a pivot language.

Approach.

We rely on a dense vector model of sentence representations, introduced by [2]. This model, called *paragraph2vec*, obtains embeddings for paragraphs, i.e. sequences of words that may range from phrases to entire documents. The *PV-DBOW* version of the model predicts the indexes of all words that occur in a sentence with a hierarchical softmax layer, thus viewing the paragraph as a bag-of-words.

We extend this model to the multilingual case, which is depicted in figure 1, where $w_n^{l_x}$ is the nth word in the sentence x in language l. A single sentence representation is instantiated for parallel sentences, from which the network predicts words that occur in the sentence in either language. This extension can easily be applied to more than 2 languages.

Word embeddings are simply defined to be the average of all sentences they occur in.

To evaluate the inter-lingual consistency, we run the cross-lingual document classification task introduced by [1] with two different datasets. In this task, a document classifier is trained using word embeddings in one language, and tested on word embeddings in another language.

Results.

We managed to create a joint semantic space with seven languages, and evaluated performance on the cross-lingual document classification in all possible combinations of two languages. Notably, the multilingual classification results are on par with the monolingual version of the task.

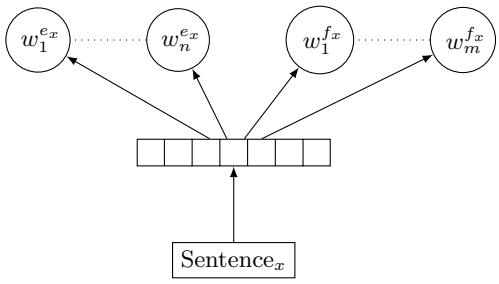

Figure 1: Extension of PV-DBOW for parallel sentences.

Conclusions.

We present a simple model to obtain cross-lingual sentence embeddings in any number of languages, that is easy and cheap to train. Word embeddings are determined in a straightforward fashion. We induce a joint space for seven languages, and show that information from one language can increase performance on document classification in another language.

1. REFERENCES

[1] A. Klementiev, I. Titov, and B. Bhattarai. Inducing crosslingual distributed representations of words. 2012.

[2] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Knowledge Discovery in Medical Forums

Erik Boertjes
TNO, The Hague

Rianne Kaptein
TNO, The Hague

Martijn Spitters
TNO, The Hague

Wessel Kraaij
TNO, The Hague

The popularity of online social networks and discussion forums has grown explosively in the last decade. Such platforms offer the opportunity to easily share information and interests with a potentially large public. For patients with a rare or common disease, social media are an effective way to exchange knowledge and support with fellow patients. Patients - being experience experts - have a lot of knowledge, experience and advice to offer about their own disease which can be useful for other patients, as well as for medical professionals. However, the huge numbers of messages, the use of spelling variants and mistakes, abbreviations and synonyms, and the often limited search functionality available on existing platforms make it hard to find the real valuable information and insights.

The goal of our prototype is to support patients as well as medial specialists in efficiently finding valuable knowledge, derive insights, and generate new hypotheses from the data available in online patient discussion forums. The prototype is built using the messages from the international GIST support group[1]. Using the Facebook Graph API we have collected over 37.000 forum messages and comments from a period of 4 years.

We address the knowledge discovery problem by combining data-driven and knowledge-driven techniques. On the one hand we apply the neural network-based `word2vec` algorithm to generate contextual word representations (word embeddings), which we use to expand the user's search queries. On the other hand we exploit existing medical domain knowledge to extract biomedical concepts from the messages.

`Word2vec`[2] implements a simple and computationally efficient way to learn word embeddings from huge data sets. In the skip-gram model, which we applied in our prototype, the training objective is to learn word embeddings which are good predictors for a word's context, i.e. the words surrounding it. The resulting vector space represents each word in terms of its contextual profile in the data. its semantic and syntactic properties. Words with a high similarity in the vector space therefore often have some strong semantic and/or syntactic relationship in the data. Such a model of word embeddings can therefore be used to discover potentially interesting relationships, simply by expanding a search query with the most similar words in the model. To illustrate this, Table 1 shows the most similar words for 'gleevec', 'kit', and 'nausea' in the vector space generated from the 37.000 GIST forum messages.

gleevec		*kit*		*nausea*	
word	score	word	score	word	score
glivec	0,77	pdgfra	0,89	fatigue	0,84
sutent	0,71	d842v	0,86	diarrhea	0,83
mg	0,68	exon	0,85	cramps	0,82
400mg	0,68	mutant	0,85	leg	0,82
gleevic	0,67	mutations	0,84	swelling	0,81
tasigna	0,66	sdh	0,82	sleeping	0,79
nexavar	0,65	fgfr	0,78	muscle	0,78

Table 1: Word2vec query expansion examples for the queries 'gleevec', 'kit', and 'nausea'.

Table 1 shows that `word2vec` returns spelling variations of the medicine gleevec (glivec, gleevic), names of related medicines (sutent, tasigna), and words related to its dosage (mg, 400mg). For the gene 'kit', word2vec returns other gene (pdgfra) and gene types (fgfr), a specific kit mutation (d842v), and the words 'mutant' and 'mutations'. Finally, the query 'nauseau' mostly yields other, often related side effects.

As a knowledge-driven discovery technique, we applied MetaMap[3], a publicly available tool for detecting biomedical concepts in text. MetaMap is part of UMLS (Unified Medical Language System)[4], which brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. In our prototype, all biomedical concepts are tagged beforehand and indexed. For the retrieval of relevant messages, we use Elasticsearch[5]. After an (expanded) search, the concepts which occur in the search result are visualized in a co-occurrence graph.

[1] `www.facebook.com/gistsupportinternational`
[2] `https://code.google.com/p/word2vec/`
[3] `http://metamap.nlm.nih.gov/`
[4] `http://www.nlm.nih.gov/research/umls`
[5] `https://www.elastic.co/`

Implementation of Specialized Product Search Features In a Large-Scale Search And Merchandising Solution

Andreas Brückner
SDL plc
Amsterdam, Netherlands
abrueckner@sdl.com

Ivan Zamanov
SDL plc
Sofia, Bulgaria
izamanov@sdl.com

Raul Leal
SDL plc
Amsterdam, Netherlands
rleal@sdl.com

ABSTRACT

Experience has shown that retrieval of products is quite different than the retrieval of documents. Although both types of retrieval share core concepts, crucial differences arise at their implementation. SDL Fredhopper is a commercial search and merchandising solution for online shops which focuses on retrieval of products and is used by more than 350 online shops worldwide.

In this demonstration we will illustrate the difference between retrieval of products and retrieval of documents by showing use cases specific to ecommerce. One important constraint is that business users are not IR specialists hence the implementation has to find a balance between the quality of information retrieved and the ease of use for the business users. Furthermore there are commercial constraints that online shops need to take into consideration when building ranking strategies (e.g. profit margins, inventory management, brand image). More concretely we will demonstrate the search and rules engines of SDL Fredhopper. These features solve the issues mentioned above by setting up search pipelines, facets, ranking strategies and manual or semi-automatic modifications to returned results.

In addition to this we will also demonstrate other ecommerce specific features of SDL Fredhopper such as language-specific requirements such us word decompounding in Germanic languages and the evaluation of search quality and processes to optimize it.

Keywords

Product search, Performance Evaluation, Quality Optimization

Self-Learning Search Suite

Manos Tsagkias
904Labs, Amsterdam, The Netherlands
manos@904labs.com

Wouter Weerkamp
904Labs, Amsterdam, The Netherlands
wouter@904labs.com

ABSTRACT

We present a demonstrator of 904Labs' self-learning search suite, which comes with a real-time dashboard. The dashboard offers insights into the current performance of the search engine, showing metrics over time, most popular queries, and failed and successful queries. Visualization of individual ranker weights also allows for in-depth analysis, as well as experimentation by disabling several rankers and observing the change in performance.

Keywords

Online learning to rank, real-time visualization

1. INTRODUCTION

An increasing number of online companies depend on search technology, from searching within apps to searching scientific literature, code documentation, and products. The importance of effective search has only recently started attracting attention outside the search industry, mainly due to positive correlations between good search results and user engagement. In the e-commerce domain, 12% of a web shop's visitors who don't find what they're looking for abandon their visit and turn to a competing shop. Large online retailers (e.g., Amazon) aim at minimizing abandonment by investing in the development of machine learning algorithms that offer visitors a personalized shopping experience. Smaller retailers cannot afford or are unaware of the importance of personalized search and recommendation reinforcing the competitive advantage of market leaders.[1] 904Labs self-learning search suite enables online retailers to improve user experience on their online properties by using state-of-the-art search and recommendation technologies.

904Labs self-learning search suite is based on latest research on search engines (cf. [1, 3]): It is a learning to rank system tightly coupled with an evaluation method, interleaving [2], which allows real-time learning and experimenting with new search algorithms. Interleaving is orders of magnitude faster than A/B testing, and minimizes the risk of harming the user group that is exposed to the new algorithm. Additionally, the system sits in between the search box and the search server of the company (e.g., Apache Solr, Elastic, Microsoft FAST, Oracle Endeca), allowing fast and easy integration in any search infrastructure.

The data flow is as follows. A visitor enters a query to the search box, the query is received by 904Labs self-learning suite, the system retrieves documents for the query and extracts features, linearly combines features, reranks documents and returns them to the visitor. When the visitor interacts with a document (e.g., via click, purchase, rating, comment) the click is sent to the system which, then, updates the feature weights. Technically, the learning method is a dual bandit gradient descent for minimizing the error on a given objective function (e.g., clickthrough rate, conversion rate). The system keeps the best weight vector so far and explores the weight space by sampling a second weight vector close to the best one. The search results from each of the two vectors are interleaved into one ranked list, and based on which document receives feedback the system decides which of the two weight vectors is best. The process repeats when a new query comes in.

At the backend, the system analyzes search user behavior in near real-time and provides useful insights to sales managers, marketing managers, and search engineers via a dashboard. The dashboard reports on search effectiveness and user engagement, and allows testing of new search algorithms/rankings on the fly (see Figure 1). This is an important feature for easily testing and deploying new search algorithms, but also in disaster scenarios where only parts of datacenters that serve features are accessible. In this scenario, the system self-heals by optimizing performance using only the available features. We are currently working on adding personalization to our system, adding historical views to the dashboard, and adding new functionality for getting insights on what queries succeed and what fail and why. These insights will help people better understand the search behavior of their visitors and may inspire new features for further boosting performance.

Acknowledgments. This work was partially supported by the University of Amsterdam under a Proof-of-Concept grant.

[1] http://venturebeat.com/2015/10/06/amazon-commands-almost-half-of-all-product-searches-and-marketers-are-ignoring-omnichannel/

2. REFERENCES

[1] K. Hofmann. *Fast and reliable online learning to rank for information retrieval*. PhD thesis, 2013.

[2] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM '08*, pages 43–52. ACM, 2008.

[3] A. Schuth, K. Hofmann, S. Whiteson, and M. de Rijke. Lerot: an online learning to rank framework. In *Living Labs for Information Retrieval Evaluation workshop at CIKM '13*, 2013.

Dutch-Belgian Information Retrieval Workshop 2015

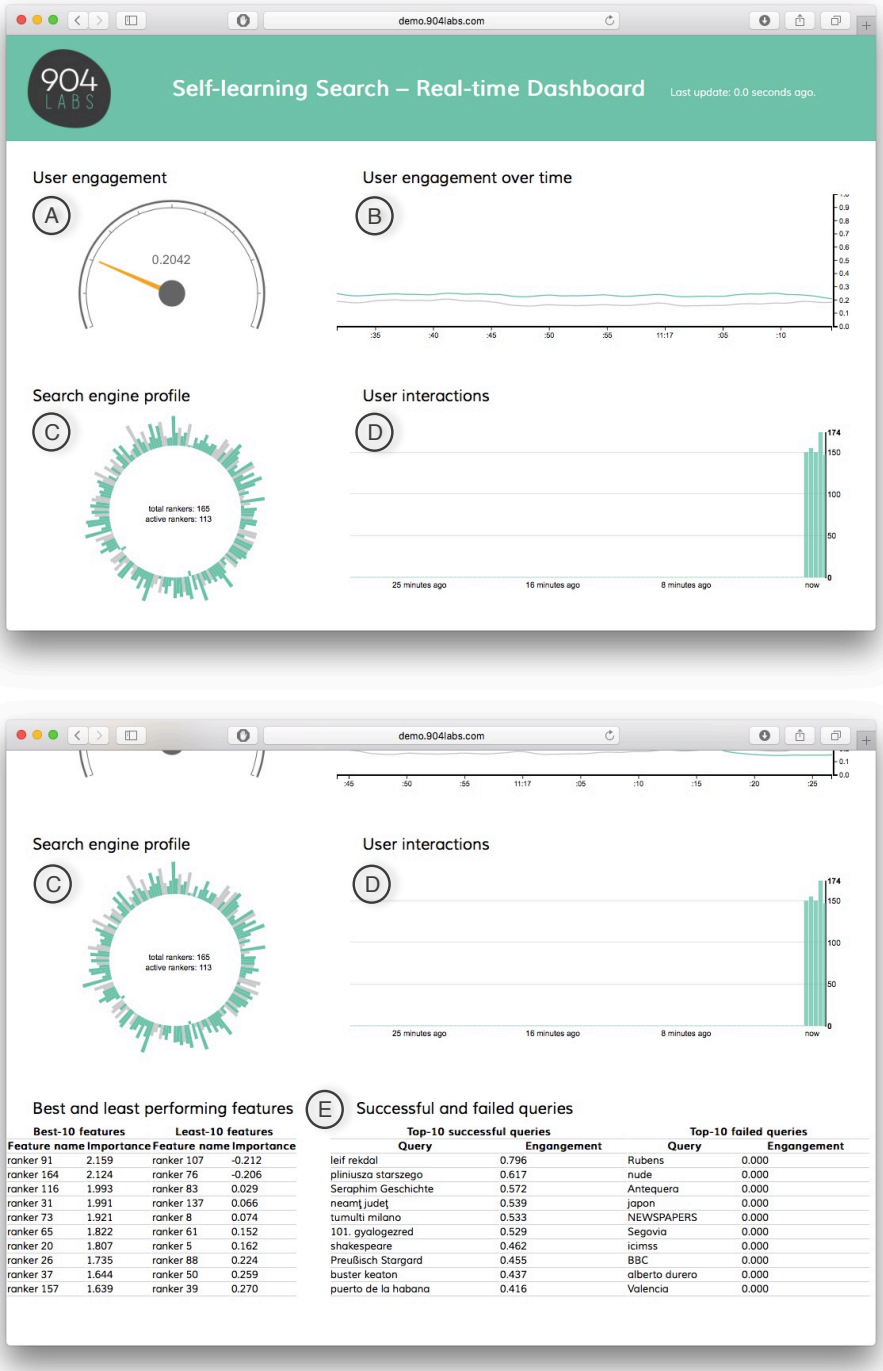

Figure 1: 904Labs self-learning search suite. (A) User engagement in normalized discounted cumulative gain (nDCG) $[0, 1]$ in near real-time, higher is better. (B) User engagement of two systems over time: 904Labs self-learning system (green line), and Apache Solr (grey line); 904Labs system improves user engagement by 38%. (C) Individual features with their importance marked as the height of the bar. Features can be activated (green)/deactivated (grey) in real-time by clicking on them; the system will try to find an optimal state for the activate features. (D) Number of queries and clicks received every 5 seconds. (E) Tables reporting on the top-10 and bottom-10 important features based on their weight, and the top-10 and bottom-10 performing queries in terms of nDCG.

Open Search

Anne Schuth
University of Amsterdam
anne.schuth@uva.nl

Abstract

It is time for a paradigm shift. Cranfield style evaluation has served us well for many years, but relevance assessments from judges are very different from what actually satisfies users. We should move to online evaluation, were we use implicit user signals to validate retrieval systems. A major issue for academics however has been the lack of a system with users.

Open Search changes this. Open Search opens up real search engines with real users for research. Open Search allows researchers to expose their retrieval system to real, unsuspecting users with real information needs that can really be satisfied.

Retrieving Research Trends in Twitter

Amna Dridi[1]

Faculty of Computer Science
Free University of Bozen-Bolzano I-39100, Italy
Amna.Dridi@inf.unibz.it

Abstract. Twitter has already become the subject of scientfic studies where it is considered as mean for academic and scientific (scholarly) communication. For instance, scientists and researchers are increasingly using social media, mainly Twitter, to discover new research opportunities, discuss research with collegues and dissiminate research information [2]. Furthermore, at scientific conferences, Twitter is often used as a backchannel to share notes and resources, and for discussion about topics [4–6]. In addition, Twitter can serve as a personal archive of information that one once found worth sharing and would like to access later on, for instance, through the use of URLs in tweets [3].
The question is how to focus on building a knowledge repository of scientific discussion in Twitter to retrieve *research trends*? The challenge in Twitter is complex due to the lack of explicit mechanisms to tell research trends from simple scientific content. For these reasons, we would adress the task of retrieving the most relevant *research trends* in scientific communication repository.

References

1. M. Stankovic, M. Rowe and Ph. Laublet. Mapping tweets to conference talks: a goldmine for semantics. In Proceedings of the 3rd International Workshop on Social Data on the Web (SDoW2010) Workshop at the 9th International Semantic Web Conference (ISWC2010) - ISWC 2010 Workshops (2010).
2. K. Holmberg and M. Thelwall. Disciplinary differences in Twitter scholarly communication. Scientometrics. November 2014, Volume 101, Issue 2, pp 1027–1042 (2014).
3. M. Mahrt, K. Weller and I. Peters. Twitter in Scholarly Communication. Chapter 30, Book: Twitter in Academia. pp 399-410 (2014).
4. C. Ross, M. Terras, C. Warwick and A. Welsh. Enabled backchannel: conference Twitter use by digital humanists. Journal of Documentation. 2010, Volume 24, No. 3, pp 183–195 (2010).
5. J. Letierce, A. Passant, J. Breslin and S. Decker. Understanding how Twitter is used to spread scientific message. In proceedings of the WebSci10 (2010).
6. K. Weller, E. Droge and C. Puschmann. Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences. Making Sense of Microposts (#MSM2011), page 1–12. (May 2011).

The Big Data Fad

Jeroen Bulters
jeroen@bulte.rs
Head of Development & Innovation
Internetbureau Holder

Abstract

In recent years "Big Data" is increasingly used for everything regarding data processing. This includes Information Retrieval being considered - in commercial circles - as a Big Data concept. Since definitions of "Big Data" are known in academic circles; recent initiatives like the Big Data Alliance are unsupportive of the clear definition of Big Data resulting in "impedance mismatch" between commercial parties and suppliers of IR/Data Science/Big Data services which might be harmfull to all involved communities.

This lightning talk is not academic in any way; born out of pure frustration with customers demanding Big Data experts when analysing small amounts of data; easily stored on consumer grade storage media from 1994. Purpose of this lightning talk is to spark discussion within the IR community and conduct a small data experiment (i.e. a poll) amongst people considered experts on IR.

Information gathered during the experiment will be shared among contributors.

Exact Match in IR

Arjen de Vries
Radboud University
arjen@acm.org

Abstract

We have become so influenced by the probabilistic ranking principle that virtually all research in IR deals with approaches that rank results. In this lightning talk I will argue that simple exact string match, provided it is sufficiently fast, serves a variety of user needs traditionally not served well, by just listing snippets with exact matches - without further ranking. I will demonstrate this hands-on using a special data structure based on the suffix array, using examples over the 400 years of scanned newspapers in the KB corpus and a collection of datasets released as open data.