Discovery and Promotion of Subtopic Level High Quality Domains for Programming Queries in Web Search

Arpita Das
Microsoft, India
arpda@microosft.com

Saurabh Shrivastava
Microsoft, India
sauras@microsoft.com

Prateek Agrawal
Microsoft, India
pragraw@microsoft.com

Sandeep Sahoo
Microsoft, India
sasaho@microsoft.com

Manoj Chinnakotla
Microsoft, India
manojc@microsoft.com

ABSTRACT

With the advancement of technology in modern era, a significant portion of the web referred to as developer segment serves to satisfy the programming related information need of the users. User satisfaction in this segment not only depends on the relevance of the retrieved pages but also on the domains that these pages belong to. We aim to discover sub-topic level associations of the domains and queries. We propose a supervised deep neural network based approach using the click-through data of a commercial web search engine to discover and promote the domains which provide high quality and expert level content for a query intent. Experiments show that our domain specific ranker performs significantly well, both qualitatively as well as quantitatively, on real-world coding query sets when compared with standard web ranking baseline. This paper further demonstrates how associating domains with query intents results in the formation of overlapping domain clusters where domains in each cluster represent a topical space of query intent(s).

CCS CONCEPTS

• **Information systems → Page and site ranking**;

KEYWORDS

domain preference, web search, user behavior

ACM Reference Format:
Arpita Das, Saurabh Shrivastava, Prateek Agrawal, Sandeep Sahoo, and Manoj Chinnakotla. 2017. Discovery and Promotion of Subtopic Level High Qual-ity Domains for Programming Queries in Web Search. In Proceedings of the first International Workshop on LEARning Next gEneration Rankers, Amsterdam, October 1, 2017 (LEARNER 2017), 5 pages.

1 INTRODUCTION

With the increase in the number of technologies and coding infrastructures, developers are becoming more and more dependent on the web. A coding query may have various intents ranging from learning basics of a programming language to debugging a code snippet. The most relevant search result for a coding query is dependent on how much the result satisfies the query intent. For example, if the query is about a particular function in a programming language, the developer will prefer a small description of the function and an example code snippet, however, if the query is about an error code he is probably looking for ways to debug it. Promoting the domain serving the correct intent will drastically improve the search engine result page(SERP).

The entire web can be grouped into intersecting clusters of domains where every cluster represents a latent topic space satisfying some query intent(s). Given a new query, we map the query to the nearest topic cluster and promote the domains associated with that cluster. For example, the query “how to format date in c#” belong to the clusters centered around coarse topics like “c#”, “time”, “changing date format” and have domains like “*docs.microsoft.com*”, “*c-sharpcorner.com*”, “*dotnetperls.com*” associated with them.

We extracted coding queries from the click logs of the commercial search engine Microsoft Bing for the past three years(2014-2016). Over this period we observed the trend of clicks for the queries with respect to 45453 coding domains. The distribution of the clicks gathered by different domains is not uniform as shown in Table 1. The domains like “*stackoverflow.com*”and “*msdn.microsoft.com*” clearly dominate the click shares. One might argue that since clicks model user satisfaction, promoting the most clicked domains for the past year might improve the SERP. Interestingly, this is not the case because ultimately the satisfaction of user will depend upon the relevance of the result with respect to the query, therefore in the domain front also, it only makes sense to promote the domain that satisfies the user intent. For the query “connecting database in azure”, from authority perspective one can assume that a developer will prefer documents from “*msdn.microsoft.com*” or “*docs.microsoft.com*” but a third domain named “*dzone.com*” exists which contain specific information about databases and their connections which exactly matches the query intent. Slight promotion of the third domain will result in the satisfaction of the user. Clicks capture the high level scenario of domain preference, but we discover and promote the domains which have high sub-topic level association with the query intent.

Retrieving intent specific domain is still unexplored in the research world. However, work has been done to detect authoritative, trustworthiness etc of domains. Traditionally researchers have used link structure based approaches and supervised approaches to predict trustworthiness of a domain. Link based approaches such as PageRank, HITS, SALSA[5] uses the structure present in hypertext

LEARNER’17, October 1, 2017, Amsterdam, The Netherlands

Domain	Clicks	Domain	Clicks
stackoverflow.com	42.01%	ozgrid.com	0.10%
msdn.microsoft.com	15.23%	powershell.com	0.10%
w3schools.com	4.29%	pandas.pydata.org	0.09%
social.msdn.microsoft.com	3.31%	community.spiceworks.com	0.09%
technet.microsoft.com	2.84%	sourceforge.net	0.09%
social.technet.microsoft.com	1.71%	getbootstrap.com	0.09%
microsoft.com	1.54%	mkyong.com	0.09%
codeproject.com	1.41%	vbforums.com	0.08%
answers.microsoft.com	1.24%	webdesign.about.com	0.08%
docs.oracle.com	1.22%	blog.udemy.com	0.08%

Table 1: Distribution of clicks among the top-100 domains specific to *coding* queries. The left half shows share of the top-10 domains while the right half shows the bottom-10 (91-100) domains.

of the web pages to identify page quality. PageRank is a well known algorithm that uses link information to assign global importance scores to all pages on the web. Bianchini et al. pointed out the vulnerabilities of the link based algorithms to spamming [4]. Since, it is possible to artificially boost authority score by forming an association of highly interlinked content, content farm websites manages to get high PageRank score. Contrary to the link based approaches, supervised approaches are robust to hyperlinked structure manipulation but they are heavily dependent on gold structured labeled data. Obtaining large-scale human judged query-domain pairs is extremely challenging in terms of cost and efficiency. Click logs are assumed to be substitution for human judged data as clicks capture human behavior and feedback to queries. Chinnakotla et al., Sondhi et al. used clicked data from web to learn supervised model to establish reliability in the health segment [7, 21]. Our paper focuses to learn the signal that is a composition of reliability, authority etc and serves the exact coding intent of the user using supervision from Bing clickthrough data.

In this paper, we propose a novel deep learning based method to maximize the conditional likelihood of a clicked domain for a given query intent. We train a three layered deep convolution neural network to project query and domains into their corresponding semantic spaces. We consider the domains with minimum semantic distance from the query to be slightly promoted in the SERP. We assume that the title of the search results in SERP is semantically relevant to the query. We make this assumption because promoting a relevant domain will not make sense if the document from that domain is irrelevant. For example, if the user query is "how to lowercase in javascript", domains like "*w3schools.com*, *stackoverflow.com*, *developer.mozilla.org*" should be promoted, however, if a document with title "how to uppercase in javascript" from "*w3schools.com*" is promoted the relevance of search result is hampered. The key contributions of the papers are : 1) We learn the deep correlations between domains and query intents for the developer segment in web. 2) We perform experiments to show how the affinity for a domain changes with a slight change in intent of the query. 3) We highlight how domains in the developer segment can be clustered based on the query intents. 4) We perform qualitative and quantitative analysis of our ranker which incorporates domain signal using large scale coding query test set and compare them with standard web ranking baseline.

2 RELATED WORK

Detecting query-intent specific domains for the developer segment in web is an unexplored problem in the world of research. However, several work has been done to solve the analogous research problems of eliminating spam websites, determining domain authority, trustworthiness, bias etc in the web.

Previous work on web spam removal or establishing reliability focused mostly on unsupervised techniques for detection of link spam (that creates tightly knit community of links to affect link-based ranking algorithm) and content spam (that malaciously spam the content of web pages). Researchers worked on automatic detection of suspicious signal in the link dependencies [1, 2, 8, 11, 20, 24] and the content of web pages [18, 19]. Castillo et al. combined link-based and content-based features and used the topology of the web graph by exploiting the link dependencies among the web pages to detect spam pages [6]. Interconnections of spam farms is also exploited to combat spam pages [3, 12, 22, 23].

Establishing authority of a web page was tried using supervised approaches too. In health domain, search results can directly impact decisions related to people's health so it is highly imperative for search engines to provide reliable information. Chinnakotla et al., Gaudinat et al., Sondhi et al. employed supervised machine learning techniques to learn the notion of trustworthiness of web pages in Health domain [7, 9, 10, 21]. Also, Hassan et al., modeled web search satisfaction of users [13–15].

Ieong et al. introduced domain bias which shows a user's propensity to believe that a page is more relevant just because it comes from a particular domain [16]. They demonstrated the importance of domain preferences in web search even after factoring out position bias and relevance. This impact of the domain bias [16] motivated us to promote documents from domains satisfying the exact query intent.

3 LEARNING INTENT SPECIFIC DOMAINS

We aim to learn a signal that promote the domains which satisfy the query intent. We use a convolutional neural network model to learn non-linear relationships between a domain and a query intent. Another way of putting it is, the neural network segment the queries into a set of fine grained topics and associate most likely domains to each of the topic space. Each topic space can be considered as a representation of a set of overlapping query intent(s).

We extracted coding queries and their clicked URLs from the Bing click logs. For feature extraction, we used character trigram based word hashing [17]. We attach the delimiter "#" to a word (say "pen" -> "#pen#") and extract its letter trigrams (#pe, pen, en#). We obtained 52339 unique letter trigrams for the entire dataset of query-clicked domain pairs. We convert each word in the query and the domain to a vector of size 52339 and mark the presence of number of occurrences of each letter trigram in the word. This representation takes care of out-of-vocabulary words and words with spelling errors.

We build a convolutional neural network with three levels of alternating convolution, max pooling and rectified linear (ReLU) layers and a fully connected layer at the top. The network gives a non linear projection of the query and domain vectors in their corresponding semantic spaces. Let x be the word hashed input term vector, y is the output vector and h is the number of hidden layers used. Let, H_j represents the j^{th} intermediate layer whose weight matrix is W_j and bias term is b_j, where j = {1, 2,. . . ,h}.

$$l_j = f(W_j H_{j-1} + b_j) \tag{1}$$

where j = {2,3,...,h} and $H_1 = W_1 x$

$$y = f(W_h H_{h-1} + b_h) \tag{2}$$

where we use tanh as the activation function f. The relevance $R(d, q)$ of a domain d for a particular query q is calculated using:

$$R(d, q) = \frac{{y_d}^T y_q}{|y_d||y_q|} \tag{3}$$

We use the supervision of the click logs to create positive and negative samples for our training data. We treat queries and the clicked domains as the positive samples (d^+) and queries and combination of domains from SERP which are not clicked for the query and some randomly selected domains as negative data (d^-). We train our network with the objective to maximize the conditional likelihood of the clicked domain given the queries or to minimize the loss function in equation 4.

$$L(\Lambda) = -log \prod_{(q, d^+)} P(d^+|q) \tag{4}$$

where Λ denotes the set of parameters of our network and $P(d^+|q)$ is the posterior probability of the clicked domain given the query.

One might question if the signal is learnt from the clicked logs of a search engine then why the search engine itself does not reflect the desired behavior already. We argue that SERP of a search engine is not only dependent on clicked signal it takes other features into account too. Also, our model does not associate a domain to the particular query, it associates domain with a topical space that represent query intent(s) and that topical space is learnt from a large collection of coding queries. For example, "*docs.oracle.com*" is not associated with the query "read a file in java" but with the topics "java", "files" etc, so when a new query "write a file in java" arrives "*docs.oracle.com*" will still be promoted.

We combine our intent specific domain score with relevance score of web ranker of Bing to promote both relevant and authoritative pages. We take the top 50 results from the initial retrieval and re-rank them using a scoring function g designed to associate relevance and authority (Equation 5). Let the initial ranker assigns scores $\{s_1, s_2, . . . , s_{50}\}$ to the top 50 URLs $\{u_1, u_1, . . . , u_{50}\}$ retrieved for a query q. Let, $\{d_1, d_2, . . . , d_{50}\}$ be the corresponding domains extracted from these URLs. The new scoring function is defined as:

$$g(q, u_i, d_i) = s_i + \alpha * R(d_i, q) \tag{5}$$

where α is the factor with which we boost the domain signal. We intentionally kept it's value small to prevent irrelevant pages from preferred domains from being promoted.

4 EXPERIMENTS AND ANALYSIS

In this section, we first describe the dataset and evaluation metric used in our experiments. We also present some interesting analysis that we can infer from the results.

Dataset Details. We collected past three years of Bing click logs and extracted queries of coding intent from them. We obtain 103 million unique query-clicked domain pairs for training the neural network. We preprocess every query by lower-casing them and removing stop words from them, we preserve the special characters as they are important in coding domain. For the preprocessing of domains we lower case them and remove prefixes like 'http' ,'https' ,'www' ,'ftp' etc if present. We run our re-ranking function on a set of 20,000 new coding queries from logs of 2017. We randomly sample 400 queries from the above set where our ranking logic introduce changes in the top 10 results and consider them as the test set. We evaluate the performance of the scoring function using our domain signal on these test queries against the current Bing ranking baseline.

Evaluation Metric. As pointed out by [7], standard IR metrics such as NDCG are not suitable for evaluating domain based signal. We also wanted to obtain a whole page comparison of the baseline and treatment therefore we chose the evaluation metric "Surplus" proposed by [7]. Following the similar setting, we show the top 10 results of baseline and treatment results to a human judge in two separate tabs in a single window. The judge can give the ratings on a seven-point scale :Left Much Better, Left Better, Left Slightly Better, Neutral, Right Slightly Better, Right Better and Right Much Better. We obtained three judgments per query for all the 400 queries in the test set to abate human judgment errors. Surplus for n queries is defined as :

$$Surplus = \frac{n_W - n_L}{n_W + n_L + n_T} * 100 \tag{6}$$

where the technique scores n_W wins, n_L losses and n_T ties.

The final metric used for measurement is $Surplus_{strong}$, where strong win/losses are used, and $Surplus_{weak}$ where weak win/losses are used. A good surplus on a large query set implies that the technique is performing well with respect to the baseline.

Results and Analysis. The result of our technique with respect to the baseline is shown in Table 2. Our technique shows significant gains in weak and strong surplus over the baseline web ranker. Table 3 illustrates the qualitative analysis of our technique. For

Query set	Number of Queries	$Surplus_{strong}$	$Surplus_{weak}$
Test set	400	**1.486**	**9.807**

Table 2: The table compares the performance of our re-ranking technique with Baseline web ranker on the test set. Results marked in boldfaced indicate that the surplus was found to be statistically significant over the baseline at 95% confidence level ($\alpha < 0.0001$). W/L/T denote the number of Wins, Losses and Ties observed.

Query: Page break html		**Query: excel vba protect sheet**	
Baseline	1.cybertext.com 2.lvsys.com 3.w3schools.com	Baseline	1.support.office.com 2.msdn.microsoft.com 3.analysistabs.com
Our Technique **(strong win)**	1.w3schools.com 2.stackoverflow.com 3.msdn.microsoft.com	Our Technique **(weak win)**	1.msdn.microsoft.com 2.support.office.com 3.mrexcel.com

Table 3: This table compares the top 3 domains shown by baseline and our technique.

Query	**Top Host**	**Query**	**Top Host**
c# string	msdn.microsoft.com	oop in python	docs.python.org
c# string out of memory exception	stackoverflow.com	oop in javascript	developer.mozilla.org
c# string tutorial	tutorialspoint.com	oop in c++	tutorialspoint.com

Table 4: This table shows how a slight change in query intent changes the affinity for most relevant domain.

the query "page break in html" we are promoting "*w3schools.com*" (which caters to the query intent in topical space of "web page structuring in html") over domains like "*cybertext.com*", "*lvsys.com*" etc. For the second query "excel vba protect sheet", apart from promoting "*msdn.microsoft.com*" over "*support.office.com*", we also promote "*mrexcel.com*" (which has specialized content in excel) over "*analysistabs.com*" .

In the process of associating domains with query intents, we found that our model inherently clusters domains whose content lie in similar topic space. We show two such clusters in Figure 1. While searching for domains similar to "*stackoverflow.com*", we observe that other forums and question-answering platforms such as "*social.msdn.microsoft.com*", "*forums.asp.net*", "*answers.microsoft.com*, "*superuser.com*", etc. come up as the closest ones. Similarly, when searched for domains similar to "*w3schools.com*", domains such as "*developer.mozilla.org*", "*tizag.com*", "*webdesign.about.com*", etc., were retrieved. Interestingly, all of these domains can be associated with a common topic space catering queries around designing web pages.

Another interesting observation that we came across is how a slight modification in query can change the affinity of domains containing relevant results. In Table 4, we demonstrate the same along two verticals. The left side portrays how a small change in query intent, with the same target coding language, changes the top retrieved domain. Whereas, the right side depicts how the change in target coding language, with same developer intent, changes the top retrieved domain.

5 CONCLUSIONS

In this paper, we proposed a novel deep learning based supervised technique to promote intent specific domains in the developer

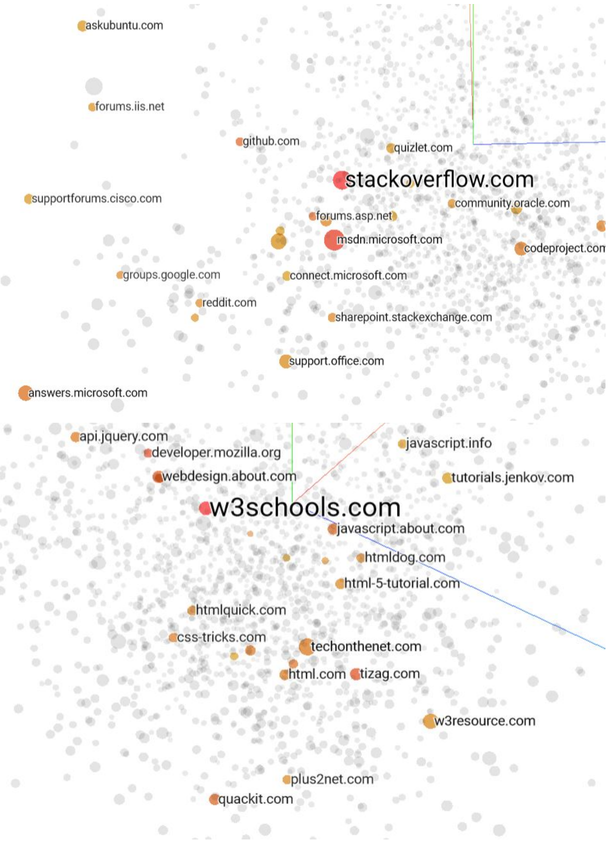

Figure 1: Examples of domain based clusters. Each cluster captures topicality of underlying query intent(s).

segment using Bing clicked logs. The evaluation metric "Surplus" proves that our method performs better than the baseline web ranking algorithm. From the experiments conducted we prove that our model segments the queries into a set of topic based clusters and associates domain with each cluster. The topicality of cluster is representation of some coarse level of query intent which the developer is looking for.

The approach proposed is re-usable and scalable in nature. Currently we have worked in the developer segment but this work can be extended to any domain. As part of future work, we plan to learn a domain signal for the entire web. Currently, we assume that the SERP contains relevant pages and slight re-ranking of pages based on domain will satisfy the users. In future, we plan to learn a signal which is a composition of query-title relevance and intent-specific domain preference and use it to re-rank results in web with more impact.

REFERENCES

[1] Brian Amento, Loren Terveen, and Will Hill. 2000. Does &Ldquo;Authority&Rdquo; Mean Quality? Predicting Expert Quality Ratings of Web Documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00).*

[2] Luca Becchetti, Carlos Castillo, Debora Donato, Stefano Leonardi, and Ricardo A Baeza-Yates. 2006. Link-Based Characterization and Detection of Web Spam.. In *AIRWeb*. 1–8.

[3] András A Benczúr, Károly Csalogány, and Tamás Sarlós. 2006. Link-based similarity search to fight web spam. In *In AIRWEB*. Citeseer.

[4] Monica Bianchini, Marco Gori, and Franco Scarselli. 2003. PageRank and Web Communities.. In *Web Intelligence*. 365–371.

[5] Sergey Brin and Lawrence Page. 2012. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks* 56, 18 (2012), 3825–3833.

[6] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. 2007. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 423–430.

[7] Manoj K Chinnakotla, Rupesh K Mehta, and Vipul Agrawal. 2014. Unsupervised Detection and Promotion of Authoritative Domains for Medical Queries in Web Search. In *11th International Conference on Natural Language Processing*. 388.

[8] André Luiz da Costa Carvalho, Paul-Alexandru Chirita, Edleno Silva De Moura, Pável Calado, and Wolfgang Nejdl. 2006. Site level noise removal for search engines. In *Proceedings of the 15th international conference on World Wide Web*. ACM, 73–82.

[9] Arnaud Gaudinat, Natalia Grabar, and Célia Boyer. 2007. Automatic retrieval of web pages with standards of ethics and trustworthiness within a medical portal: What a page name tells us. *Artificial Intelligence in Medicine* (2007), 185–189.

[10] Arnaud Gaudinat, Natalia Grabar, Célia Boyer, et al. 2007. Machine learning approach for automatic quality criteria detection of health web pages. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. IOS Press, 705.

[11] Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Link spam alliances. In *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 517–528.

[12] Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Link spam alliances. In *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 517–528.

[13] Ahmed Hassan. 2012. A Semi-supervised Approach to Modeling Web Search Satisfaction. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*.

[14] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: User Behavior As a Predictor of a Successful Search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)*.

[15] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM international conference on Conference on information &#38; knowledge management (CIKM '13)*.

[16] Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. 2012. Domain Bias in Web Search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*.

[17] Paul Mcnamee and James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information retrieval* 7, 1 (2004), 73–97.

[18] Gilad Mishne, David Carmel, Ronny Lempel, et al. 2005. Blocking Blog Spam with Language Model Disagreement.. In *AIRWeb*, Vol. 5. 1–6.

[19] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*. ACM, 83–92.

[20] Guoyang Shen, Bin Gao, Tie-Yan Liu, Guang Feng, Shiji Song, and Hang Li. 2006. Detecting link spam using temporal information. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 1049–1053.

[21] Parikshit Sondhi, VG Vinod Vydiswaran, and ChengXiang Zhai. 2012. Reliability Prediction of Webpages in the Medical Domain.. In *ECIR*, Vol. 12. Springer, 219–231.

[22] Baoning Wu and Brian D Davison. 2005. Identifying link farm spam pages. In *Special interest tracks and posters of the 14th international conference on World Wide Web*. ACM, 820–829.

[23] Baoning Wu, Vinay Goel, and Brian D Davison. 2006. Propagating Trust and Distrust to Demote Web Spam. *MTW* 190 (2006).

[24] Hui Zhang, Ashish Goel, Ramesh Govindan, Kahn Mason, and Benjamin Van Roy. 2004. Making eigenvector-based reputation systems robust to collusion. In *WAW*, Vol. 3243. Springer, 92–104.