Place Value: Word Position Shifts Vital to Search Dynamics

Rishiraj Saha Roy, Anusha Suresh and Niloy Ganguly
IIT Kharagpur, India - 721302.
{rishiraj, anusha.suresh, niloy}@cse.iitkgp.ernet.in

Monojit Choudhury
Microsoft Research India
Bangalore, India - 560080.
monojitc@microsoft.com

ABSTRACT

With fast changing information needs in today's world, it is imperative that search engines precisely understand and exploit temporal changes in Web queries. In this work, we look at shifts in *preferred positions* of segments in queries over an interval of four years. We find that such shifts can predict key changes in usage patterns, and explain the observed increase in query lengths. Our findings indicate that recording positional statistics can be vital for understanding user intent in Web search queries.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation

General Terms

Experimentation, Human Factors, Measurement

Keywords

Position shifts, Query log analysis, Query understanding

1. INTRODUCTION

With constant social and technological changes, new words are entering the lexicon and preferred usages of existing words in Web search queries are evolving quickly over time. It is crucial that search engines are able to identify and harness these temporal changes to ensure timely interpretation of queries. We investigate the change in relative positions of words and phrases (referred to as *segments* in this text) within queries – an important aspect of search dynamics that, we believe, has a lot of potential but have been overlooked in past research. To be precise, we compare the relative positions of query segments over an interval of time and find that positional shifts have ramifications on the *roles* of these segments in respective queries. Our findings indicate that search queries are undergoing substantial structural changes, one of the outcomes of which is an increase in their average length.

In particular, we extract aggregate word usage statistics (Sec. 2) from the 2006 AOL log sampled from the USA [1] and a 2010 Bing log sampled from Australia, containing $12.8M$ and $11.9M$ queries respectively, which are a good four years apart; the average lengths of distinct queries in these logs are 3.50 and 3.98 words respectively. Duplicates were retained to preserve the natural power law frequency distribution ([1] and own analysis) of queries.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. ANALYSIS APPROACH

Segments and roles: To ensure that segments are temporally meaningful, we applied a state-of-the-art unsupervised query segmentation algorithm [2] that mainly relies on query logs to learn the segments. Recently, researchers have shown that segments in Web search queries can be categorized into two classes according to their *role* in the query [5]: *Content segments (c)* that carry the main semantic content of the query (e.g., `titanic` and `machine learning`), and *Intent segments (i)* that are specified by the user to indicate their intent in the context of the content words (e.g., `movie review` and `define`). Segments that *co-occur* (appear in the same query) with a large number of distinct segments over the entire query log (i.e., have a high *co-occurrence count*), are more likely to indicate user intent [3, 4, 5]. Here we consider the top 2000 segments, when sorted in descending order of co-occurrence counts, as *intent segments*, and the rest as *content*; we will study the temporal dynamics of intent and content segments separately.

Positional statistics: Absolute positions (e.g., the 3^{rd} segment of a query) are sensitive to various factors (including query length) and thus cannot be used as meaningful statistics. Instead, we define three relative positions – *beginning (b)*, *middle (m)* and *end (e)* for a segment s in a query. From the query log, we compute the following three probabilities: $P_b(s)$, $P_m(s)$ and $P_e(s)$ – the probabilities of observing s in the beginning, middle and end of a query respectively. Thus, $P_b(s)$, $P_m(s)$ and $P_e(s)$ add up to one. If a segment is the only one in a query, its position is taken to be *b*. For every s that appears in both the logs, we computed these statistics and the occurrence probability $P_{occ}(s)$ for both years. Subsequently, we calculated their corresponding changes over time and looked for meaningful patterns.

3. OBSERVATIONS AND INSIGHTS

Our results are summarized in Table 1. Changes in positional probabilities were considered to be meaningful only if their absolute values were above the significance threshold of 0.05. We break the set of all segments into four categories as shown and identify interesting correlations between positional changes and segment roles. The **#Against** reports the number of segments that significantly violate the expected trend. As we can see, in all the four cases, the frac-

Table 1: Summary of identified trends in positional transitions for each category with real examples.

Category	#Segments	Trend	#Against	Example 1	Example 2	Example 3	Example 4
$c \rightarrow i$	445	$P_b(s) \downarrow$	52	`youtube`	`imdb`	`xbox`	`ps3`
$i \rightarrow c$	576	$P_e(s) \downarrow$	95	`yellow pages`	`white pages`	`motels`	`clipart`
$c \leftrightarrow c$	9293	$P_b(s) \downarrow$ or $P_e(s) \downarrow$	83	`solar system`	`eiffel tower`	`epilepsy`	`washing machines`
$i \leftrightarrow i$	1253	$P_b(s) \uparrow$ or $P_e(s) \uparrow$	19	`how to`	`define`	`download`	`reviews`

tion of such segments is extremely low, supporting our observations mentioned below. We now examine each of these categories separately discuss relevant insights obtained.

Content then, intent now: Most of these segments were used in a standalone fashion as navigational queries in order to issue internal searches (`imdb, youtube`), or as informational searches (`xbox, ps3`). With growing popularity of these units (as verified through their increased $P_{occ}(s)$) and the ability of the search engine to address the information needs of users directly, queries like `titanic imdb` and `halo 3 xbox` have become more common, effectively converting their roles to intent specifiers. Since most of these segments now appear only to the right of content segments, they show marked drops in $P_b(s)$ and consequent rise in $P_m(s)$ or $P_e(s)$.

Intent then, content now: 576 segments were identified as content in 2010 that were previously labeled as intent (like `yellow pages` and `clipart`). These segments were mostly popular intent identifiers in 2006 which have gradually fallen out of favour over the next few years (as verified through their decreased $P_{occ}(s)$). In 2010, these units were mostly issued as standalone segments, possibly as esoteric interests of specific users. Earlier, they appeared mostly to the right of content segments. Thus, they show significant drops in $P_e(s)$, and associated rise in $P_b(s)$. Positional statistics can hence be viewed as predictors of certain units becoming obsolete in the near future, for reasons attributed to technological change or emergence of more influential entities. An interesting case was observed during *error analysis* in this zone. Segments like `what to do`, `cheat codes` and `official site` were tagged as c in 2010, which are obviously errors on part of the labeling algorithm. This happens due to the decreased co-occurrence counts of these segments in 2010, which in turn is caused by their decreased $P_{occ}(s)$. However, they do not violate the observed positional trends for standard intent segments (increase in $P_b(s)$ or $P_e(s)$, Row 4 in Table 1). Thus, we recommend that such role labeling decisions must be taken considering both positional trends as well as traditional co-occurrence counts.

Content then, content now: Segments that have remained content are mostly entities (`eiffel tower`) or classes (`washing machines`) of some kind. While content segments were popularly issued as standalone queries or with a single qualifier in 2006, increased specificity of user needs have added intent words to the left (`meaning of`, `pics of`) or right (`map, movie review`) (generally not both) of the former class of segments. So they show noticeable drops in their $P_b(s)$ or $P_e(s)$. This phenomenon contributes to the increased mean query length over the years.

Intent then, intent now: The traditional intent words, that have retained their usage, also show definitive patterns indicating their stabilization in the query structure. Such words prefer to be at the ends (`reviews`) or the beginnings (`how to`) of queries (except for items like `and` and `between`, that always prefer to be in the middle), and show observable rise in the associated probabilities. However, the generally preferred positions of intent segments remains the end (`sony stocks latest updates` preferred over `latest updates on sony stocks`). This can be explained by a user model of *query formulation* where the content part is conceived first, followed by the specification of the associated requirements [5]. Several intent segments show significant gains in $P_{occ}(s)$, reflecting the relative abundance with which users add qualifiers to their queries now, as compared to the scenario four years back.

4. CONCLUSIONS AND FUTURE WORK

In this work, we have highlighted the importance of recording positional statistics of segments in search queries. We have identified simple rules concerning relative positions of segments, and used them to predict possible changes in usage. Consistent use of words like `imdb` and `wikipedia` beside content segments have forced search engine designers to formulate *user-guided rules*, and today one can almost always expect to find links from IMDB and Wikipedia in the first rank for relevant queries. Identifying query intent is one of the most important challenges faced by search systems today. Since most of our findings are directly related to intent phrases, it is a worthwhile suggestion that alongside other indicators, system designers must keep track of relative postions for segments to foresee probable role changes. We also note that while `titanic` was a more commonly expected query in the the past, it is not surprising to come across `titanic movie review imdb` today – this *stacking* of intent segments has been found to be the dominant factor towards increased query length. This is a work in progress and needs more rigorous experimentation with a larger number of temporally separated query logs from diverse geographical regions.

5. REFERENCES

[1] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale '06*, 2006.

[2] R. Saha Roy, N. Ganguly, M. Choudhury, and S. Laxman. An IR-based Evaluation Framework for Web Search Query Segmentation. In *SIGIR '12*, pages 881–890, 2012.

[3] X. Yin and S. Shah. Building taxonomy of Web search intents for name entity queries. In *WWW '10*, pages 1001–1010, 2010.

[4] X. Yin, W. Tan, X. Li, and Y.-C. Tu. Automatic extraction of clickable structured Web contents for name entity queries. In *WWW '10*, pages 991–1000, 2010.

[5] H. Yu and F. Ren. Role-explicit query identification and intent role annotation. In *CIKM '12*, pages 1163–1172, 2012.