Computing Science

The effect of corruption on search engine query prediction

Charlie F. Egan

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: November 15, 2015

CS4040 Report

Department of Computing Science **November 2015**
University of Aberdeen
King's College
Aberdeen AB24 3UE

The effect of corruption on search engine query prediction

Charlie F. Egan

Department of Computing Science
University of Aberdeen

November 15, 2015

Abstract: Search engines have are able to learn from user behavoir to improve their results. Rather than relying on a traditional spelling correction algorithms, user query patterns can be used to train the system and provide more accurate predictions. This report investigates how increasingly corrupted input effects the prediction accuracy of the system and well as how each in a number of top search engines compare in their ability to predict intended input. Related work and concepts are also discussed. It was found that search engines differ significantly in their prediction accuracy and that there is no relation between the word length performance.

1 Introduction

Search engines offering query predictions are the research interest of the project. Search engines offer predictions to overcome corruption in user queries, where corruption is the combined effect of one or more typographical errors. While prediction performance is better than that of traditional spelling correction algorithms SOURCE the detailed processes behind predictions are not publicized. On a fundamental level, failed user queries followed by the user's own corrections are used to build a list of corrupted inputs for a desired result SOURCE.

The project aim is to evaluate the variation in the prediction accuracy in responses, for increasingly corrupted terms, queried on a number of common search engines. Prediction accuracy is capability of a search engine suggest or adopt the intended term, given a corrupted version. The effect of query length in characters, for a given rate of corruption, is also investigated as part of the project.

Information about search engine prediction accuracy for corrupted queries would be useful for those making mechanical use of search engines for spelling correction SOURCE. A ranked list of search engines by prediction accuracy could also be of interest to those with poor spelling or dexterity from conditions such as Dyslexia or Parkinson's disease. Also of interest would be the relationship between daily users and the performance of query prediction as a feature.

Sections 2-7 cover: spelling correction background as a task in natural language processing, the research questions, the design of the comparison experiments, the results, their discussion and conclusion respectively.

2 Background and Related Work

A review of related work [2, 1, 3].

Guide length: 500 words.

Further explain terms and problem and reading you have done - even if not papers, give a focus for my paper, ”others have looked into...”, only need 2-3 relevant papers that are well discussed.

Computerized spelling correction has been an active area of research since YEAR SOURCE. Early approaches, as well as more simplistic implementations today such as those found browsers, offer predications for strings with typographical errors by calculating *edit* or *Levenshtein* distance. The edit distance of two strings is the number of *edit operations* required to transform one string to another SOURCE 200 PAGE. In 1966 Levenshtein described a model for such transformations SOURCE 1996. *Levenshtein distance* has been the basis for many such implementations since (Error Model for Noisy Channel Spelling - section 2 SOURCE if required).

Not only is calculating edit distances for an string against terms large dictionaries is computationally expensive (SOURCE An adaptive spell checker based on ps3m) these implementations are limited by the words as described by the dictionaries. There is no account taken of the relative key positions or that (SOME OTHER THING ABOUT TYPOS) SOURCE. For example, *Facebokk* is more likely than *Facebozk* to be a corrupted query for *Facebook* though they have the same Edit and Levenshtein distances. Also it is not possible to make phonetic comparisons between strings, such as between *'an introduction to **sea** programming'* and the expected result.

3 Research question

Given the problem context (Section 1) and background (Section 2), you should now be in a position to present what you have investigated. **Pose this as a question.**

Then you should present your approach to addressing this question.

Guide length: 500 words.

”intro why asking questions”, ”the questions are as follows”, ”At which point does”, and as a second ”is there a significant difference between the search engines”

”in order to answer this”, ”overview”, explanation of tools

”the datasets...”, explain how generated, how controlled, list all the variables in the data (suggested in the design section)

How the results are recorded and on what they will be compared (precision recall), or use accuracy? Need to explain why, accuracy might be appropriate for my experiment. This is a summary of the following section. High level parameters for the tools and generally how the subject are compared.

4 Experimental Design

What are your hypotheses? How are you going to test them? What is your target population? What are your datasets; i.e. your sample of the target population. What are the dependent and independent variables?

Guide length: 500 words.

first state the null hyps, one per question it seems, ”significantly different” is the phrase to use it seems. can also state the alternatives if they are not clear.

detail test data and give justification for decisions

”To test these hyps [what happens in the experiement]”, ”this was then repeated for each...”

talk about the datasets that will be used. talk about the variables ()

5 Results

Present the results. A good way to organise this is via subsections for each hypothesis you tested. Include graphs of results , tests of significance, etc. If you have negative results, include them. A negative results is just as informative and useful as a positive one, sometimes more so.

Guide length: 500 words.

table of the results in summary, same sentence ref a graph of the table.

the results suggest that all decreased in accuracy as the number of corruptions increased. Also x did better then y - quote values to support / highlight things in the table

then do a section for each comparison and hypothesis / quesion 5.1 5.2 - say was was and was not significant, in response to the question. Quote those p values!

Segment the section based on the number of questions, accuracy vs speed etc. for each quote the values of the statistical tests.

6 Discussion

What do the results say? What have you learned from the experiments? Have you identified a correlation between variables, or causation? What are the limitations of what you’ve done? What further experiments might be of benefit?

Guide length: 400 words.

results summary, state the relationships at a high level, make references to the values of the tests mean stdev etc.

limitations of the study. limited in data and tools compared. talk about any bias, are my typos real for instance? likely not.

How to do it better next time, make better representation of the population, big pg on what was missed and how this might be an issue.

7 Conclusion

What have you done and why? What have you shown through your experiments?

Guide length: 100 words.

tiny, 1 pg on the work done, why done ”id the best” / ”test the difference”, results summary, best worst wrt hyps,

I have conducted an experiment to test...

What are the implications of the results? What do they mean at a higher level?

References

[1] C. Burnett, T. J. Norman, and K. Sycara. Stereotypical trust and bias in dynamic multi-agent systems. *ACM Transactions on Intelligent Systems and Technology*, 4(2):26:1–26:22, 2013.

[2] Ian J. Taylor & Andrew Harrison. *From P2P and Grids to Services on the Web*. Springer, 2 edition, 2009.

[3] Wikipedia. Peer-to-peer networks. `http://en.wikipedia.org/wiki/Peer-to-peer`. Accessed: 1st May, 2015.