Towards Correlating Search on Google and Asking on Stack Overflow

Chunyang Chen, Zhenchang Xing
Background

- Stack Overflow
  - 12m questions, 19m answers, 5.6m users;
  - A huge treasure with knowledge covering most parts in Software Engineering;
  - Many research works had been carried out to mine Stack Overflow
Can Stack Overflow represent the interest of developers around the world?

Many research works had been carried out to mining Stack Overflow, but no one considered this question.

We expect to demonstrate the representativeness of Stack Overflow so that to lay the solid foundation for further mining in Stack Overflow.
Are there relationships between the queries developers use in search engines and the questions developers ask in Q&A sites?

Google is the dominating search engine and most developers use Google
- Its search log can definitely indicates developers’ interest
- Google Trends provides the statistics of a search item that people use as query to search Google

Once we can demonstrate that correlation exists between Google Trends and Stack Overflow, Stack Overflow is representative.
Select technical terms (i.e., tags in Stack Overflow):
- E.g., c#, visual-studio, ios, jquery, android-layout (no ambiguity)

Collect frequent search and asking technical terms
- i.e., collect frequent co-occurring keywords of terms mentioned above

Build search and asking trends for technical terms
- Collect time-series data of term usage frequency both in Stack Overflow (300 weeks) and Google (574 weeks) and normalize them to 0~1 for comparison.
Given 185 technical terms, 65% of their co-occurring terms in Google and Stack Overflow overlap.
Patterns of search and asking trends

- There are six trend patterns

- Distribution of term number in different trend patterns

Similar number in each pattern
We adopt cross correlation to compute the Pearson correlation coefficient and delay between two trends for each term.

**Algorithm 1: Cross correlation of search and asking trends**

**Input:** Search trend $T_s$ and Asking trend $T_a$ of a technical term
**Output:** maxCorrelation, delay
maxCorrelation = -1;
for $w$ in $-239 : 35$ do
    $T_{s-seg} \leftarrow T_s.getSegment(w, w + T_a.len)$;
    $r \leftarrow \text{pearson}(T_{s-seg}, T_a)$;
    if $corr > \text{maxCorrelation}$ then
        maxCorrelation $\leftarrow r$;
        delay $\leftarrow w$;
    end
end
The newer technology, the higher correlation between two sources.

At the launch of SO, no obvious correlation
We also explore the delay between trends.

At the launch of SO, it is always behind Google.

The newer technology, the shorter delay between two sources.
Some general terms also have different specific version number

- E.g., visual-studio: vs2008, vs2010, vs2012
- python: python-2.7, python-3.x

High correlation for specific terms (vs2008, python-2.7), low correlation for general terms (visual-studio)
Replacement process
- New technology is replacing old ones;
- High correlation also exists between aggregated asking and search trend (e.g., asp.net-mvc).
Conclusion

- There is correlation between asking in Stack Overflow and searching in Google especially those new technologies.

- The results ensure the representativeness of Stack Overflow as a software repository for further research.

Thanks for listening

Chen Chunyang