



Query suggestion using synonymy and Excite logs

Kalyan Boggavarapu

CSE 497

Lehigh University

What is Query Suggestion?

- Possible alternate query provided by the system to the user
- Many search engines provide the search query alternatives
- Like Google, AltaVista, etc.,

Uses

- The user may not be familiar with the vocabulary in the topic he is searching for

- Eg: User: University Bethlehem

Sugg: **Lehigh University Bethlehem**

User: multinational corps

Sugg: **multinational corporations**

User: Nirvana

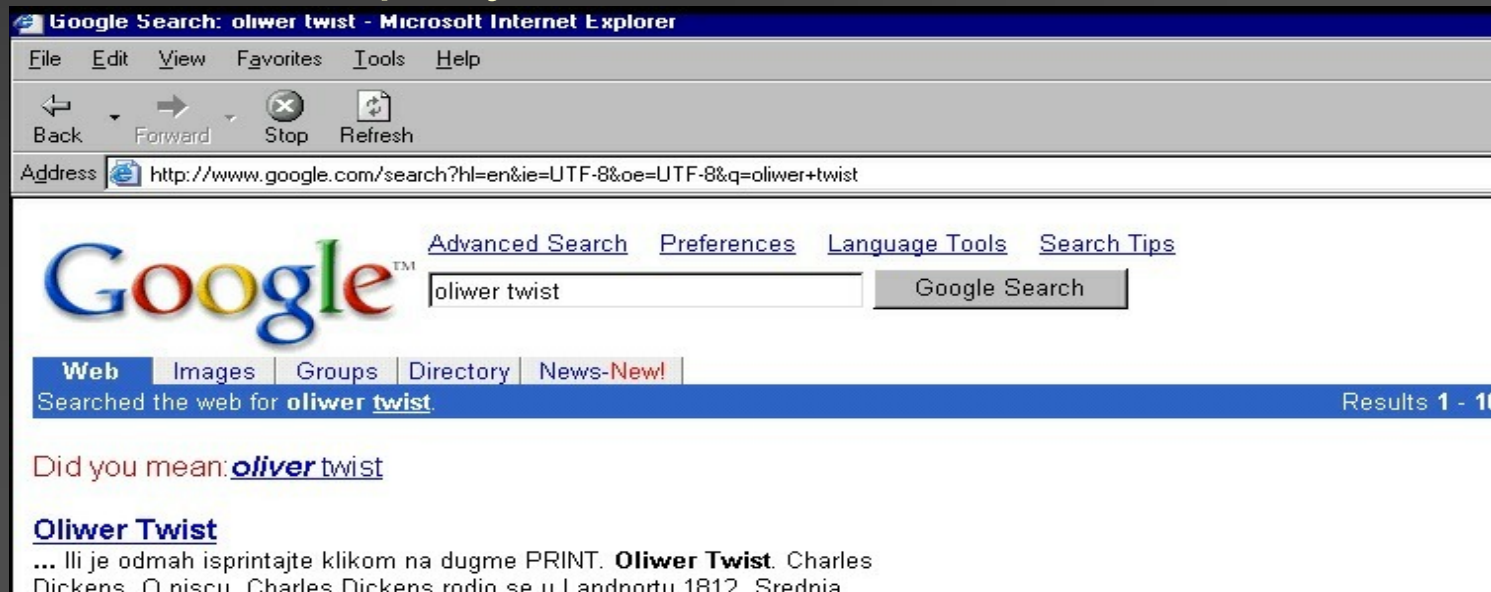
Sugg: **Nirvana album, attaining Nirvana**

User: stergis rally

Sugg: **sturgis rally**

Query Suggestion

- Types of query suggestions:
 - Spell Check
 - Suggest synonyms
 - Suggest a popular word , a little variant of the current word in the query



The screenshot shows a Microsoft Internet Explorer browser window with the title "Google Search: oliver twist". The address bar contains the URL "http://www.google.com/search?hl=en&ie=UTF-8&oe=UTF-8&q=oliver+twist". The search bar contains the text "oliver twist" and a "Google Search" button. Below the search bar, there are links for "Advanced Search", "Preferences", "Language Tools", and "Search Tips". A navigation bar includes "Web", "Images", "Groups", "Directory", and "News-New!". Below the navigation bar, it says "Searched the web for **oliver twist**". A suggestion box displays "Did you mean: **oliver** twist". The search results section is titled "Oliver Twist" and begins with the text "... Ili je odmah isprintajte klikom na dugme PRINT. **Oliver Twist**. Charles Dickens. O niscu Charles Dickens rodio se u Landportu 1812. Srednja

Are there any other approaches for finding similarity between words?

How is spell check done?

- Where do we find the alternatives?
 - Use a online dictionary
 - Use a thesaurus in memory
 - Using logs:
 - Popular queries which are similar to the query
- How is similarity measured between words?
 - One method:
 - Calculating Distance between words
 - Eg: Roger is similar to Rogeer
 - parameters: SameNumLetters=5
 - LettersSamePosition=4

Our Approach

- Query suggestion:
 - Spell check
 - Synonym word substitution
 - Popular word substitution
 - Rank the alternate queries

Stage 1

Should Stop-words be removed from the Query Logs?

Query

Where is Lehigh coltege

Remove Stop Words

lehigh coltege

Spell check using Aspell

lehigh college

Get the Synonyms

lehigh college, lehigh university
lehigh school

Stage 2

Search the query logs for the frequency
of the alternate queries

Get the number of hits from the web
For the alternate queries

Rank the alternate queries

Present them to the user

lehigh college 56
lehigh university 100
lehigh school 17

lehigh university 100
lehigh college 56
lehigh school 17

Refinement and Searching

- Remove the stopwords
 - Stopwords: stopwords are common words which do not have searching significance
 - Like then, always, here etc....
 - Query logs: Excite query logs of about 2.5 Million queries, collected over a day
 - They provide:
 - Time, Query-Id, set of results seen, Query
 - Number of results returned to the user
- Eg: 0930 023569 20 "Philadelphia Eagles"

Perform Spell check

- Spell check by Aspell

- Aspell is a improved version of Unix Ispell.
- Eg: might correct some spell check like
 - Ealges --→ Eagles
- uses Metaphone Algorithm for correcting the Surnames

Find Synonyms

- Get the synonyms for each word
 - Wordnet is used to get the synonyms for each word
 - Wordnet is a online thesaurus
 - Words are arranged in hierarchical structures of senses
 - Each sense contain a set of words
 - So each synonym would a sense numbers
 - This number provides **distance** between a sense and the input word
 - <http://www.cogsci.princeton.edu/cgi-bin/webwn1.7.1>

Questions?
