Sara Stolbach  
Natural Language Processing HW 3  
Due: November 29, 2007  
Search Engine Improvement Proposal

Abstract

People often use the wrong name or fact to refer to an entity. This is very common among products, people, and locations. I propose a solution to aiding the search of such users on the new search engine Guggle. Please note that I interned at Google this past summer. This works is my own and does not refer to an idea I heard or worked on during my time there.

Introduction

There are often scenarios when a person wants to search for some word or phrase but they are not sure what it is they are looking for or they have mistaken the actual fact. This is particularly common amongst places, people, products, and media and their authors. For example, suppose a user wishes to learn about the location of the Google headquarters but mistakenly think that it is in “Mountain Valley”, instead of the actual location of “Mountain View” (Perhaps confusing the exact name with Microsoft’s headquarters Silicon Valley). A method is proposed that will automatically suggest to the person “Perhaps you meant Mountain View”

Related Work - Proof Idea Doesn’t Exist

Currently search engines “rely” on companies to have multiple definitions. For example, a website would tag Grand Central Station as Grand Central Station and Grand Central Terminal causing it to appear in the search results via either query. This is cumbersome for the website and it does not signify that it is the best result.

Suppose a person is interested in finding out information about the Civil War (see appendix A), but are mistakenly searching for the revolutionary war. Many of the top search engines have been examined including Google, Ask.com, Yahoo, and MSN to examine their results. Consider the initial query: revolutionary war. Satisfying results are certainly not achieved so the user modifies the query: revolutionary war slaves. The user would still not receive satisfying results on any of the top websites mentioned [1,2,3,4] (see appendix B,C,D,E) and therefore the user modifies the query yet again by adding more terms: revolutionary war slaves north south.
After the third try, Google [1] (see appendix I) and Ask.com [2] (see appendix H) mention the civil war within the first page, and Google has it as the top answer. The user could then realize their mistake and change their query to civil war to produce better results. MSN [3] (see appendix E) and Yahoo [4] (see appendix G) do not even show results related to the civil war on the top page meaning more refining would need to be done before the user could notice their error. In the best situation, the user will have to make 4 queries prior to receiving the results they wish. 3 queries to realize their mistake and 1 correct query. There were two opportunities to help the user achieve their goal earlier.

The first opportunity was after the user entered ‘revolutionary war slaves’. A query of ‘civil war slaves’ displays 1,970,000 (see appendix K) results on Google [1] whereas ‘revolutionary war slaves’ displays 1,830,000 (see appendix B) results. The difference is even more evident when “civil war” and “revolutionary war” are placed in double quotes, having a difference of 1,900,000 (see appendix L) to 884,000 (see appendix M) respectively, which leads me to believe that many of the results in the query with revolutionary war actually mention another war, probably the civil war. A method will be proposed in the next section that should recognize this difference and suggest to the user “Perhaps you meant the civil war”.

The second opportunity is at the point where the user recognizes their mistake; when the correct fact is listed in the top results. For example, on Google [1], the first result points to the Wikipedia entry for the civil war (see appendix I). This information can be used to suggest to the user “Perhaps you meant the civil war”, which will cause them to have to type one less query and will help them feel more confident in the correction of their mistake. The method proposed in the next section should aid the user in correcting their mistake at this point.

Finding the Intended Query

The possible mistakes can be extremely common or less frequent. Popular mistakes are those in which the top result is often the right answer, due to either statistical relevance (mistake is common) or website tagging (website can tag their site with multiple terms including the wrong ones to improve ranking). For example, suppose the user queries for DisneyLand, Florida instead of the proper query, Disney World, Florida. This is an extremely common mistake and the top result is still the Disney World site (see appendix N). These mistakes are not of any concern since they already return a useful result, however this could be recognized and given to the users as an option to easily fix their query by selecting “Perhaps you meant Disney World, Florida”. It will allow the search engine to the aid the user and not require the website to manually enter the various possible tags.

The less common mistakes can be broken up into three categories and each must be dealt with differently. The
first category is a regular query. The second category is a query with the incorrect portion in double quotes; for example, “Disney World” Florida. The third and final category is a query with the incorrect portion within a word; for example, querying for headbook instead of facebook (mainly applies when one word query, it is essentially a subset of the main problem). This paper will be focusing on the general or regular mistakes but briefly mention some ideas for the other mistakes within the future work section.

This paper proposes an approach to finding the user’s intended query based upon the current query by examining portions of the query with other words. The other words are obtained by taking the best features within the query results of the portions of the query. The algorithm is described below, and then the revolutionary/civil war example from the previous section is used to show how it will work. It is important to note that it is assumed prior to the calling of this algorithm that other techniques have already been called upon such as stemming, removal of stop words from the query, and spelling corrections.

---

**Other Fact Suggestion Algorithm**(x, tu, nf)

1. Keep list, Q, of user queries during current session
2. For each query that satisfies some threshold of uncertainty, tu
   3. Add query to Q
4. Let w be the number of words in a query
5. Divide the current query into w portions of size w-1
6. Let f be a unique list of features (use a hash)
7. For each n-gram
   8. Eliminate if it is in the list of previous queries, Q
9. Compute the best features based on the top x query results
10. Add the top features into f based on a threshold, nf
11. Rank the features in f in descending order
12. Suggest the best ranked feature

---

The previous section mentioned that the first change could have been made when the user entered the query revolutionary war slaves. Assume that the query satisfies the threshold of uncertainty mentioned in step 2. The threshold of uncertainty is satisfied when there is enough proof that the user is unsure of the best query due to factors such as modifying the query and not visiting any pages mentioned in the results. The query can be broken up into n-grams of size w-1. This would give three queries: revolutionary war, war slaves, and revolutionary slaves. revolutionary war could be eliminated because that was a previous query which clearly did not satisfy the
user. Features must be computed based on the results for each n-gram and then merged into one list to find the best feature overall.

The only non-trivial step within the outer loop is computing the best features for each n-gram. All the other steps will run in almost constant time because the number of n-grams should be fairly small because most people tend to use small queries. Even in the case of large queries the runtime will be fast because as the query size gets larger the number of results will get smaller. Computing the best features can be achieved a number of ways. One possible method would be to get a subset of words using a window size surrounding the query words in each of the top $x$ results; for example, the text shown as part of the result. The title can also be used. Testing would need to be done on the subset of words to find the best feature extraction method. For example, the chi-square test or document frequency (with a list of stop words) can be used on the subset to create a ranking of important features. It also may be worthwhile to use part of speech to only examine nouns (or more specifically proper nouns) since most mistakes occur on those words. This method is clearly space efficient, but can take time depending on the precise method used and the number of words. (Computing the document frequency for each word would take $O(n)$ and then sorting the list would take $O(n \log n)$ where $n$ is the number of words in the subset.). One possible solution is to store the best suggestion after it is made, especially if the user is satisfied and if it is a common error, because the runtime will be constant from then on and that is more important than the minimal amount of space it will require.

The total runtime for the algorithm is $O(qnk)$ where $q$ is the number of queries that must be evaluated before the user is satisfied, $n$ is the number of n-grams and $k$ is the time to run step 9. If the algorithm is successful, $q$ should be small. The limitations of this algorithm are clear. It is possible that finding a suggestion can take a significant amount of time due to $k$. It is important to minimize the number of words that need to be examined. There is almost certainly a method for accomplishing this quickly, efficiently and successfully, but experiments would need to be run to find the best solution. Once this obstacle has been overcome, it can easily be used for any query on the web. If little additional storage is needed, the cost of implementing the algorithm should be fairly minimal.

In this algorithm, precision is more important than recall because it is better to be correct than to show a meaningless suggestion. There are a number of thresholds in place that should improve precision. The main threshold that will improve precision is only showing a suggestion if user dissatisfaction is evident. This will improve user satisfaction because the user does not want to be bothered by suggestions stating they have their facts incorrect.
An insight to the hope of the ranking can be visualized by looking at just the first page of the n-gram queries for *revolutionary war slaves*. Using Google [1], the query results for *revolutionary slaves* refer to the American Revolution 4/10 times, and the Haitian Revolution 4/10 times on the first page (see appendix O). The query results for *war slaves* refer to the civil war 8/10 times on the first page (see appendix J). Ranking these results it is clear that civil war was the most probable intended query and it can be suggested to the user “Perhaps you meant the *civil war*”.

**Future Work**

In addition to the general case, the mistaken fact can be within a word. This is more difficult than the general case. The first step would be to divide the word into parts. Then, it would be useful to experiment with Wordnet [5] because it may be more likely in these cases that the person used the wrong synonym, such as *headbook* instead of *facebook*. The rest of the algorithm would be similar to the general case.

Mistaken facts can also occur within double quotes. One option would be to simply run the algorithm by ignoring them. Some search engines, such as Ask.com [2] do get rid of double quotes if no results show up. However, it may be more probable that the mistake is within the quoted section. If that is the case, the algorithm can be simplified to only look at the n-grams where the entire phrase within the quotes occurs.

**Conclusion**

In conclusion, giving the user suggestions to mistaken facts is a step forward in the computer doing the ‘thinking’ for the user in a time when it is most beneficial to them - when they don’t even know they have made a mistake at all. This addition to the Guggle search engine will improve its quality, precision, and user satisfaction causing Guggle to rise as the top search engine available today.

**References**

[1] www.google.com
Appendix

A. Google Search – Civil War

B. Google Search – Revolutionary War Slaves
1. **Africans in America | Part 2 | The Revolutionary War**
   The British actively recruited slaves belonging to Patriot masters and, ... Boston King was an escaped slave who joined up with the Loyalists and later ... www.pbs.org/wgbh/asiapart2/2main4.html - 18k - Cached

2. **American Revolutionary War | Wikipedia, the free encyclopedia**
   African-Americans—slave and free —served on both sides during the war; ... have recruited more slaves and Native Americans to fight the war, but this would ... Quick Links: Civil Rights before 1776 - Slaves, militias, and mercenaries - African Americans and Native Americans on.wikipedia.org/wiki/American_Revolutionary_War - 13k - Cached

3. **Slaves and indentured in Revolutionary War**
   The Community Environmental Legal Defense Fund ... as the body or person of a Negro, man or woman, who is sold as a legal Slave ... celdf.org/DemocracySchool/2h4-2h5/2h5Default.aspx - 26k - Cached

4. **Revolutionary War | Definition and Much More from Answers.com**
   The war shocked slavery severely, and thousands of former slaves also departed with the British. ... slave and free —served on both sides during the war. ... www.answers.com/topic/american-revolutionary-war - 265k - Cached

5. **Revolutionary War | The Home Front**
   ... a “home front” in the Revolutionary War is difficult because so much of ... years of the war took place in the South. many slaves escaped to British lines. ... memory.loc.gov/ammem/amlav/html/revhomeфр.htm - 13k - Cached

---

C. Yahoo Search – Revolutionary War Slaves

D. Ask.com Search – Revolutionary War Slaves
E. MSN Search – Revolutionary War Slaves

F. MSN Search – Revolutionary War Slaves North South
1. Southern theater of the American Revolutionary War. Wikipedia, the free encyclopedia. 
   ... the north, focused on campaigns around the cities of Boston, New York, and Philadelphia. ... 2 British campaign in the South. The Loyalist question. ... 
   Quick Links: Early operations, 1775-1776 - First attack on Charleston - British campaign in the South 
   en.wikipedia.org/wiki/Southern_theater_of_the_American_Revolutionary_War ... - 67k - Cached

2. American Revolutionary War. Wikipedia, the free encyclopedia. 
   ... to now Canada to the north, Florida to the south, and the Mississippi River to the west. ... 
   Revolutionary War | National liberation movements | Wars of ... 
   Quick Links: Combatants before 1778 - Arrows, missiles, and mercenaries - African-Americans and Native Americans 
   en.wikipedia.org/wiki/American_Revolutionary_War ... - 135k - Cached

3. The American Revolutionary War. 
   ... to now Canada to the north, Florida to the south, and the Mississippi River to the west. ... more 
   slaves and Native Americans to fight the war, but this ... 
   www.myrevolutionarywar.com/index.htm ... - 22k - Cached

4. African American Participation in the Revolutionary War. 
   In the South, slaves grew tobacco in Maryland, Virginia, and North Carolina, and ... African 
   Americans, slave and free, also worked in a wide variety of ... 
   www.americaneverywar.homestead.com/files/BLACKS.HTM ... - 14k - Cached

   ... duly returned north, nominating Lord Cornwallis as his successor in the south. ... War. The 
   colonial wars were the background to the Revolutionary War. ... 
   www.answers.com/topic/american-revolutionary-war ... - 200k - Cached

---

G. Yahoo Search – Revolutionary War Slaves North South

---

H. Ask.com Search – Revolutionary War Slaves North South
I. Google Search – Revolutionary War Slaves North South

American Civil War – Wikipedia, the free encyclopedia
Both North and South assumed that slavery could not expand or it would wither and die. ... en.wikipedia.org/wiki/American_Civil_War - 234k - Cached - Similar pages - Note this

American Revolution – Wikipedia, the free encyclopedia
It was the first fighting of the American Revolutionary War. ... gradual abolition occurred in the North, but slavery expanded in the South and racial ... en.wikipedia.org/wiki/American_Revolution - 112k - Cached - Similar pages - Note this

African American Participation in the Revolutionary War
This portrait (left) of an unidentified Revolutionary War sailor was painted in all ... www.americansrevwar.homestead.com/files/BLACKS.HTML - 14k - Cached - Similar pages - Note this

The Civil War
In the South, most slaves did not hear of the proclamation for months. But the purpose of the Civil War had now changed. The North was not only fighting to ... www.pbs.org/gwwh/laa/part4/14p2007.html - 21k - Cached - Similar pages - Note this

Slavery in the North
Slavery in the North never approached the numbers of the South ... practical economics; and the fact that the Revolutionary War had broken the Northern ... www.slaverynorth.com - 15k - Cached - Similar pages - Note this
J. Google Search – War Slaves

K. Google Search - Civil War Slaves

L. Google Search – “Civil War” Slaves
M. Google Search – “Revolutionary War” Slaves

N. Google Search – DisneyLand Florida

O. Google Search - revolutionary slaves