Finding the Geographic Focus of Web-Pages

Einat Amitay Nadav Har'El Ron Sivan Aya Soffer
IBM Haifa Research Lab
Haifa 31905, Israel
{einat,nyh,rsivan,ayas}@il.ibm.com

1. INTRODUCTION

The problem of finding geographic names mentioned in text documents (such as Web Pages) has been attacked repeatedly by a number of researchers; Our own paper appearing in this conference, [1], gives one approach and refers to others. Clearly, simple text search would yield relatively low precision, as searches for "London" intending to find the English capital would find unrelated mentions of London, Ontario (a Canadian city with population of 350,000), of the author "Jack London", and so on. Limiting the search to explicit mentions of "London England" will drastically lower recall, as many pages mention London, the city in England, without writing "London England" explicitly. Therefore, searching for geographic names requires some sort of a *disambiguation* stage, figuring out (using contextual clues, and other techniques) to which places the text is referring to.

But, once we determined the correct meaning of every geographical name mentioned in the page, we would also like a way to separate the wheat from the chaff — to decide which geographic mentions are incidental, and which constitute the actual *focus* of the page, i.e., a place (or very small number of places) that the page mainly discusses. Knowing this focus might be useful, for example, if the user wants to search for pages about California, rather than finding the multitude of pages that mention in passing some city in California or pages that list all the states of the union.

This presentation discusses our focus-finding algorithm implemented within the framework of our Web-a-Where system ([1]). The talk will also survey other people's approaches, and suggest ideas for future research. We also present a way to test the precision and recall of a geographic focus algorithm (as compared to decisions made by humans) by using massive amounts of free existing data, rather than resorting to slow and expensive manual inspection of the algorithm's decisions.

2. FOCUS SPECIFICATION

When designing a focus finding algorithm, one of the key issues is how it will specify its answer. How will we specify which region is in focus?

One approach is to use a given geographic hierarchy, also known as a *gazetteer*. The gazetteer lists known geographic entities in a hierarchical manner, specifying for each place its parent (containing region). Our gazetteer (described in [1]) contains a hierarchy of continents, countries, states (for some countries) and cities (those with more than 5,000 inhabitants), for a total of 40,000 places around the world. A focus-finding algorithm that uses a geographic hierarchy takes as input a page where certain phrases point to gazetteer nodes (e.g., a "Paris" string in the page points to the Paris France node, and a "London" points to the London England node), and as a result gives one or several hierarchy nodes, in this example we might return the Europe node.

Because of the simplicity of the specification of the results, our focus-finding algorithm uses gazetteer nodes for both its input and its output.

Superficially, this approach is limited by not being able to recognize as foci regions which are not in the tree-like gazetteer, such as *The Tri-state area*, *Southern California* or *The Middle East*. But many algorithms, including our own, actually do not limit the gazetteer to having a tree structure — each node can have several parents, rather than one (e.g., *France* can be considered to be in *Western Europe*, and in the *European Union*, in addition to it being in *Europe*).

3. OUR FOCUS-FINDING ALGORITHM

Our basic premise is that if several cities from the same region are mentioned, this might mean that this region is the focus. For example, a page mentioning San Francisco (Calif.), Los Angeles (Calif.) and San Diego (Calif.) can be said to be about California. A page mentioning San Jose (Calif.), Chicago (Ill.) and Louisiana can be said to be about the United States. A page that is predominantly about the United States with a single mention of Paris France can still be said to be only about the US. Repeated mentions of the same place count: A page mentioning the state of California five times is probably just as likely to be about California as a page mentioning five different cities in California.

Sometimes we cannot say that a page has only one focus. For example, two different countries might be repeatedly mentioned in some news story. In such cases we will want to list several geographic regions as foci. However, we must still try to coalesce many places into one region before declaring foci, so that a page that lists the 50 states of the United States will not be said to have 50 separate foci, but rather one focus — the United States.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.

The other extreme should be avoided as well: if a small region is the real focus of a page, we should not unnecessarily report a larger region. It is all too easy, but not very productive, to report several continents as being the "focus".

The focus-finding algorithm assumes that all geographic names have already been disambiguated correctly. When the disambiguation algorithm makes a bad guess, it should give it a low confidence score. In finding the focus, we should take these confidence scores into account, giving higher weight to information coming from locations with higher confidence.

The algorithm proceeds as follows. Each geographic mention, disambiguated into a hierarchy node (e.g., `Paris/-France/Europe`) with confidence $p \in [0, 1]$, adds a certain score ($s = p^2$ in our implementation) to the importance of this place in the page. It further adds lower scores to the enclosing hierarchy nodes: sd to `France/Europe` and sd^2 to `Europe`, where $0 < d < 1$ (0.7 in our implementation) is a decay factor. We sum up the scores contributed by all places in the page, and then sort the hierarchies by score. We loop over them from from highest to lowest, stopping at the low threshold (0.9 in our implementation), or if sufficiently many foci were already found (in our case, 4). We skip nodes that cover or are covered by one already selected as focus. Otherwise we add this node to the list of foci.

The aforementioned weights and thresholds are based on some experimentation, but they are by no means optimal. Additional research is needed to discover the optimal values.

The reason that places contribute less score to their enclosing regions (the d decay factor) is that this allows the more specific place to "win" if it is the only place mentioned in this region, while permitting the region to be chosen as focus if several different places in it are mentioned with no emphasis on any.

An example will make the algorithm clearer: A certain page contained four mentions of `Orlando/Florida` (confidence 0.5), three `Texas` (0.75), eight `Fort Worth/Texas` (0.75), three `Dallas/Texas` (0.75), one `Garland/Texas` (0.75), and one `Iraq` (0.5). A human that was asked to judge what is the geographical focus of this page responded with "It's about Texas and perhaps also Orlando". Indeed, that page comes from the "Orlando Weekly" site, in a forum titled "Just a look at The Texas Local Music Scene...". Our scoring algorithm gave the following scores:

```
6.41 Texas/United States/North America
4.97 United States/North America
4.50 Fort Worth/Texas/United States/North America
3.48 North America
1.68 Dallas/Texas/United States/North America
1.00 Orlando/Florida/United States/North America
0.70 Florida/United States/North America
0.56 Garland/Texas/United States/North America
0.25 Iraq/Asia
0.17 Asia
```

The algorithm proceeds to go over this sorted list from the top. Texas got the top score (because several separate cities — Fort Worth, Dallas and Garland contributed to it, even though each city contributed more to its own score) and is chosen as a focus. The next highest scorer, the United States, already covers Texas so it is dropped: it doesn't make sense to say that both Texas and something that covers Texas are in focus. The next scorer, Fort Worth, is covered by Texas and is dropped for the same reason, as are North America and Dallas which follow it in the list. We then get to Orlando/Florida, which does not cover the existing focus of Texas nor is it covered by it, and is taken as a second focus. The remaining scores (e.g., for Iraq/Asia) are below the importance threshold (0.9) and are ignored. This page therefore ends up with two foci: Texas and Orlando, with Texas being the first (stronger) focus.

92% correct up to country level			*8% incorrect country*	
38% Precise match	30% Correct state or city	24% Correct country	4% Correct continent	4% Continent wrong

Table 1: Comparison of Web-a-Where-determined focus to human-determined one (ODP)

4. EVALUATION

We evaluated the focus algorithm by comparing its decisions to those of human editors. The Open Directory Project (`http://dmoz.org`) is the largest (4 million pages) human-edited hierarchical directory of the Web, and is maintained by a vast community of volunteer editors. It's "Regional" section is devoted to English-language pages with a coherent geographic focus (e.g., sites about a place, or about a company located in a certain city). This availability of over one million real Web-pages pre-tagged with their geographic focus allowed us to automatically test the focus algorithm on a very large sample of Web-pages, much bigger than we could afford to manually tag ourselves. We ran Web-a-Where on a random sample of over 20,000 Web-pages from the ODP's Regional section that were larger than 3k, and compared the foci it reported to those listed in the ODP index. The results of this comparison are given in table 1.

Several points are due noting when interpreting these results: First, the focus algorithm relies on individual geographic mentions to have been disambiguated correctly. Therefore, this test evaluates not just the focus algorithm, but also the underlying disambiguation algorithm. Second, the ODP editors based many of their focus decisions on information not available to our algorithm, such as images and sub-pages (27% of the pages did not have a single geographic name). Third, the determination of a page's focus is more subjective than the meaning of a single geographic name. It's not always clear if a page is about "England" or about the "United Kingdom", for example. Finally, the ODP's hierarchy contains many metro areas, regions, and towns too tiny to appear in our gazetteer. Accordingly, we cannot always expect an exact match between the focus reported by our tagger and the one listed in the ODP. Instead, we need to define what qualifies a "good enough" match. Therefore, Table 1 divides these matches into several quality types.

Given these difficulties, our algorithm did quite well. It found a page focus in 75% of the pages with one or more geotags. Of the pages with focus, in 91% the focus had the correct country, in 65% the focus matched up to the 3rd hierarchy level (state or city), and in 38% the focus matched the ODP listing precisely.

5. REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging Web Content. In *proceedings of ACM SIGIR*, Sheffield, UK, July 2004.