Mining Topological Relations from the Web

Steven Schockaert
Ghent University, Gent, Belgium
Steven.Schockaert@UGent.be

Philip D. Smart, Alia I. Abdelmoty, Christopher B. Jones
Cardiff University, Cardiff, Wales, UK
{P.Smart,A.I.Abdelmoty,C.B.Jones}@cs.cardiff.ac.uk

Abstract

Topological relations between geographic regions are of interest in many applications. When the exact boundaries of regions are not available, such relations can be established by analysing natural language information from web documents. In particular, we demonstrate how redundancy–based techniques can be used to acquire containment and adjacency relations, and how fuzzy spatial reasoning can be employed to maintain the consistency of the resulting knowledge base.

1. Introduction

Qualitative spatial relations, and topological relations such as containment, overlap and adjacency in particular, are paramount in geographic information retrieval (GIR) systems [3, 4]. Users of a GIR system may be interested, for instance, in information about holiday resorts *in* Southern Europe which are *adjacent* to a beach. When the boundaries of the regions involved are known to the system, topological relations can, in principle, be checked by straightforward geometrical computation. In some situations, however, very little information may be available about these boundaries, e.g., gazetteers typically only provide a single point (centroid) to approximate the location of geographic regions. To assess which topological relations hold between two regions, additional knowledge is therefore required. Moreover, many of the regions people refer to in everyday communication do not have official boundaries; as a consequence, the topological relations between them are inherently ill–defined and subjective.

A promising idea to obtain information about unknown and/or ill–defined topological relations is to analyse web documents that refer to the regions of interest. Unfortunately, automatically extracting relations from natural language text is a challenging problem [1], which is in the spatial domain even further complicated by the sparseness, in web documents, of explicit mentions of spatial relations. To cope with some of these problems, we propose to employ heuristic techniques, initially valuing high recall over high precision. To arrive at reliable conclusions, the extracted information is subsequently filtered using a fuzzy spatial reasoner. Experimental results indicate that in this way, a reasonably comprehensive and accurate knowledge base of topological relations can be obtained. As a case study, we focus in this paper on the neighbourhoods of Cardiff in Wales, UK.

2. Related Work

GIR systems are quickly gaining importance, and a wide array of computational techniques to support it have already been explored. A central theme is recognising occurrences of place names in texts, and determining the corresponding geographical coordinates (e.g., [10]). A geographical scope is thus assigned to every web page, which is compared in the retrieval process with a geographic constraint specified explicitly by the user, or implicitly by the user context. Several researchers have specifically focused on the ambiguity of place names (e.g., there are over 80 places called Springfield in the US Census Bureau gazetteer). To decide which place is referred to in a text, often the proximity to other places mentioned in the same text are taken into account (e.g., [5]): if all places in a text are French, then a reference to Paris in this text is likely to refer to the capital of France. Another technique for place name disambiguation, which is related to some of the filtering steps we apply below, is based on co-occurrence of place names [6]. Another relevant line of research attempts to augment the geographical information found in gazetteers by analysing web pages. For example, [11] studies techniques to find cognitively significant landmarks in cities, while [2] is concerned with approximating the boundaries of imprecise geographic regions.

3. Containment Relations

Containment relations are omnipresent in web documents, but unfortunately, they are most frequently implicit.

By far the richest source of containment relations are addresses of the following kind

```
Monthermer Road 67, Cathays, Cardiff,
Wales, UK
```

Typically, such addresses mention increasingly larger regions; e.g., from the address above we can derive that Monthermer Road 67 is located in Cathays, which is a part of Cardiff, which is in Wales, which is in the UK. Web documents contain a wealth of addresses, although parsing addresses is often complicated by occurrences of HTML tags. For example, in the following address, different constituents are placed on different lines[1].

```
Aberdare Hall<br /> Corbett Road<br />
Cathays<br /> Cardiff <br /> Wales
```

Therefore, we do not only need to look at commas when parsing addresses, but also at HTML fragments such as
. In addition to addresses, implicit containment relations are sometimes found in URLs, e.g.[2]:

```
action="/gallery/Europe/United_Kingdom/
        Wales/Cardiff/Roath/"
```

Again, we witness a series of containment relations, although now the largest region is mentioned first (Europe), followed each time by a smaller subregion. Finally, implicit containment relations are often formulated in ad hoc ways[3]:

```
Cardiff - Cathays - Alexandra Gardens
```

The only constant in all these examples is that regions are ordered from largest to smallest, or from smallest to largest, and between each region name, the same separating string occurs. Knowing that region R_2 is part of region R_1, we can therefore expect to find parts R_3 of R_2 by looking for occurrences of the pattern $R_3\#R_2\#R_1$ or $R_1\#R_2\#R_3$, where # is a recurring, but otherwise arbitrary string (e.g.,
). This heuristic can be useful to identify those regions from a set $\mathcal{R}$ of region names that are located in R_2, but, being a heuristic, it is bound to fail occasionally. Moreover, if we have no prior knowledge at all about the regions that can possibly be located in R_2, it is not always easy to correctly parse the name of R_3. Consider, for example, the following HTML fragment[4]:

```
Icon courtesy of St. Martin Church,
Roath, Cardiff, Wales
```

How do we decide that "St. Martin Church" is part of Roath, rather than "Icon courtesy of St. Martin Church"? One possibility is to change the patterns to $\#R_3\#R_2\#R_1$ and $R_1\#R_2\#R_3\#$", in which case fewer matches will be found, but for each match, we should be able to correctly identify R_3. Another, more heuristic solution is to rely on capitalisation, which will result in more containment relations, but with a lower accuracy.

As a first experiment, we tried to identify the names of the neighbourhoods of Cardiff. To this end, we retrieved about 4000 relevant documents using Google and Yahoo!. Next, we scanned these web documents for matches of the patterns $\#R_3\#R_2\#R_1$ and $R_1\#R_2\#R_3\#$", where R_2 is Cardiff and R_1 is one of South Glamorgan, Glamorgan, Wales or UK. The strings matching R_3 are possibly neighbourhoods of Cardiff, but they need, in fact, not be place names at all. To filter out matches that do not correspond to places, two techniques are used. The first technique is based on some manually defined rules (e.g., the first and last word need to be capitalised). The second technique is based on the observation that if R_3 is a place name, then Google should at least find some results for the queries "located in R_3", "located in the R_3", "situated in R_3" or "situated in the R_3". If the four queries together yield less than 5 results, R_3 is filtered. Most of the remaining matches actually correspond to Cardiff neighbourhoods. However, some notable errors still occur, including place names that are not contained in Cardiff (e.g., Wales, Swansea, Edinburgh, Europe), as well as strings that do not correspond to place names at all (e.g., Women, The, Shopping, Bar), thus necessitating a further filtering step.

For both types of errors, the names found are not likely to co–occur with the term "Cardiff" very often. To assess whether R is likely to be a region in Cardiff, we therefore use Google to estimate the number of web documents q_1 containing R, as well the number of web documents q_2 containing both R and Cardiff. If $\frac{q_2}{q_1} > 0.75$, we can be quite confident that R is indeed in Cardiff. The converse, however, is not necessarily true. For example, while "City Centre" corresponds to a neighbourhood in Cardiff, most documents containing "City Centre" will not contain "Cardiff". Another way of identifying place names in Cardiff is by geocoding addresses involving these names and looking at the spatial distribution of the resulting coordinates. Let $\mathcal{P}$ be the set of points (coordinates/addresses) that are thus found for a region R. Let p_0 be the most central point of $\mathcal{P}$ (medoid), r_1 the median of $d(p_0, p)$ over all p in $\mathcal{P}$, and r_2 the median deviation, i.e., the median of the value $|d(p_0, p) - r_1|$ over all p in $\mathcal{P}$. Below, we assume that $d(p, q)$ corresponds to the distance between p and q in kilometers. The more the points of $\mathcal{P}$ are clustered together over a relatively small area, the higher the chance that R is indeed a neighbourhood. Assuming that $r_1 + r_2$ is a reasonable ap-

[1] http://www.cf.ac.uk/locations/locationsaz/index.html, accessed March 4, 2008.

[2] http://www.trekearth.com/gallery/Europe/United_Kingdom/Wales/Cardiff/, accessed March 4, 2008.

[3] http://screenandsound.llgc.org.uk/cronfa/placesindex.php?index=C, accessed March 4, 2008.

[4] http://www.stmartins-charlotte.org/, accessed March 4, 2008.

Table 1. Neighbourhoods in Cardiff after additional filtering.

cardiff university	llanrumney	roath park
roath	radyr	splott
cardiff bay	birchgrove	old st. mellons
rumney	penylan	high street arcade
canton	ely	st.-mellons
cathays	grangetown	docks
llanishen	rhiwbina	mermaid quay
cardiff	heath	old-st.-mellons
llandaff	penarth	llandaff-north
llanedeyrn	riverside	taffs-well
whitchurch	fairwater	st.-fagans
pentwyn	cowbridge road east	st fagans
tongwynlais	cathays park	lakeside
st. mellons	gabalfa	old st mellons
heath park	lisvane	central cardiff
atlantic wharf	leckwith	caerau
llandaff north	pontprennau	butetown
cardiff gate	barry	cardiff castle
culverhouse cross	marshfield	llantwit major
cardiff gate business park	the hayes	adamsdown
taffs well	pontcanna	pontypridd
st mellons	rhoose	danescourt
thornhill	university hospital of wales	

proximation of the radius of R, this suggests that the likelihood of R being a neighbourhood is inversely proportional to $(r_1 + r_2)^2$. On the other hand, the more addresses found in Cardiff referring to the region name R, i.e., the higher the chance that R is a neighbourhood, suggesting that the likelihood that R is a neighbourhood is proportional to $|\mathcal{P}|$. Specifically, if $r_1 + r_2 > 0$ and $\frac{|\mathcal{P}|}{(r_1+r_2)^2} > 1$, we assume that R is a neighbourhood, but again the converse does not hold, i.e., there may be neighbourhoods in Cardiff for which no addresses are found. This leads to the following additional filter step: if either $\frac{q_2}{q_1} > 0.75$ or $\frac{|\mathcal{P}|}{(r_1+r_2)^2} > 1$, R is considered to be a neighbourhood name. In other words, if none of the two techniques finds evidence that R is a neighbourhood in Cardiff, we assume that R is either outside Cardiff or not a place name. The resulting neighbourhood names (ignoring case) are provided in Table 1. Out of the 68 neighbourhoods, seven are duplicate entries (e.g., st. mellons and st mellons), four can be considered vernacular or colloquial (cardiff bay, central cardiff, mermaid quay, atlantic wharf), ten are not considered to be neighbourhoods of cardiff (cardiff, cariff castle, high street arcade, university hospital wales, cowbridge road east, cardiff university, cardiff gate business park, barry, pontypridd, rhoose) and five are either close to, but not within Cardiff (penarth), or are areas within cardiff that are not recognised neighbourhoods (cathays park, roath park, heath park, the hayes).

4. Adjacency Relations

If explicit mentions of containment relations in texts are already rare, this holds even more for adjacency relations. One exception is when people state that something is located in the border zone between two neighbourhoods. From the following sentence, for example, we can establish that Cathays and Roath are adjacent neighbourhoods[5]:

> 4 Double Bedroom house located on the border of Cathays and Roath.

Although this kind of information is often expressed, many variations on the exact phrasing are possible, e.g.[6]:

> Small 1 bedroom flat in the Cathays/Roath area.

In general, people use adjacent neighbourhoods often in the same context. To assess the likelihood that two regions R_1 and R_2 are adjacent, we therefore count the number of times we find occurrences of "R_1 / R_2", "R_1 & R_2" and "R_1 and R_2".

Unfortunately, this technique requires a prohibitively high number of search engine requests. For example, considering the 68 neighbourhoods from Table 1, at least $68 \times 67 \times 3 = 13668$ search engine requests would be needed (assuming we submit the queries 'R_1 / R_2", "R_1 & R_2" and "R_1 and R_2" to Google or Yahoo!). Therefore, rather than considering all pairs of regions R_1 and R_2, an initial filtering step is performed. As for containment relations, two complementary techniques are used: one based on co–occurrence and one based on addresses. For the first technique, we use the documents that were already collected for R_1 and R_2 to find containment relations. In these documents, we count the number of times f that R_1 and R_2 occur within 100 characters of each other. If $f > 5$, we apply the method described above to assess the likelihood that R_1 and R_2 are adjacent. For the second heuristic, let $\mathcal{P}^1$ and $\mathcal{P}^2$ be the coordinates of locations that were found to be in R_1 and R_2 respectively, and let p_0^1 and p_0^2 be the corresponding medoids. Furthermore, let r_1^1 and r_1^1 be the median distances from p_0^1 and p_0^2, and let r_2^1 and r_2^2 be the median deviations from these median distances (w.r.t. $\mathcal{P}^1$ and $\mathcal{P}^2$ resp.). If a sufficiently high number of coordinates was found for R_1 and R_2, the distance between p_0^1 and p_0^2 should be small compared to the values of r_1^1, r_2^1, r_1^2, r_2^2. In particular, we assume that $r_1^1 + 2r_2^1$ and $r_1^2 + 2r_2^2$ are reasonable approximations of the radius of R_1 and R_2 respectively. Therefore, if $d(p_0^1, p_0^2) < 0.2 + r_1^1 + r_1^2 + 2(r_2^1 + r_2^2)$ (assuming $d(p_0^1, p_0^2)$ is the distance in kilometer between p_0^1 and p_0^2), we consider R_1 and R_2 as a pair of possibly adjacent neighbourhoods and apply the aforementioned method.

[5] `http://2let2students.co.uk/CMS2/index.php?option=com_hotproperty&task=view&id=299&Itemid=114`, accessed February 13, 2008.

[6] `http://www.nestoria.co.uk/cathays/flat/rent`, accessed February 13, 2008.

5. Fuzzy Spatial Reasoning

The techniques introduced above yield a high number of containment and adjacency relations, whose accuracy, however, is rather limited. To cope with this, we employ a fuzzy spatial reasoner to detect and repair inconsistencies among the various topological relations. In other words, it is hoped that incorrect relations result in inconsistencies, and can thus be identified. Note that in [9], a similar strategy was pursued for improving the accuracy of extracted information in the temporal domain. Furthermore note that, in addition to improving accuracy, the fuzzy spatial reasoner is useful to identify new relations by applying (fuzzy) compositional inferences. The exact algorithm we use is similar in spirit to the algorithm from [9]; we omit the details.

Rather than using a classical spatial reasoner, we utilise fuzzy spatial reasoning to cope with spatial relations that are inherently ill–defined. In particular, due to the vagueness of the exact boundaries of many city neighbourhoods, it is often unclear what the "right" topological relation is, and different people can hold different opinions about this. One example is the relationship between Cardiff Bay and Butetown, where some people would consider Cardiff Bay as a part of Butetown and others as being adjacent to it (see below). It is important to differentiate this scenario with occurrences of clearly wrong topological relations in the knowledge base. To represent topological information, we therefore use a fuzzy region connection calculus [8], in which topological relations are defined as fuzzy relations. Below, we will specifically use the fuzzy relations EC, P and DC, modelling adjacency, containment and disjointness respectively. For example, for two regions a and b, $EC(a, b)$ is the degree in $[0, 1]$ to which a and b can be considered adjacent. If $EC(a, b) = 1$, a and b are perfectly adjacent, while a and b are not adjacent at all if $EC(a, b) = 0$. On the other hand, if $EC(a, b) = 0.5$, for instance, it is not entirely clear whether a and b are adjacent or not. We refer to [7] for an operational semantics of these membership degrees, as well as algorithms for reasoning about these fuzzy spatial relations.

Initially, adjacency and containment relations are interpreted as being true to degree 1. For example, if we derive using the method from Section 3 that Cardiff Bay is a part of Butetown, we add $P(\textit{Cardiff Bay},\textit{Butetown}) = 1$ to the knowledge base. In addition to containment and adjacency relations, we add a number of relations of the form $DC(R_1, R_2) = 1$, based on available coordinates. Let $\mathcal{P}^1$ and $\mathcal{P}^2$ be sets of coordinates of places in R_1 and R_2, and let p_0^1, p_0^2, r_1^1, r_1^2, r_2^1 and r_2^2 be defined from $\mathcal{P}^1$ and $\mathcal{P}^2$ as before. Assuming that $r_1^1 + 2r_2^1$ and $r_1^2 + 2r_2^2$ are good estimations of the radius of R_1 and R_2, if $d(p_0^1, p_0^2)$ is much larger than $r_1^1 + r_1^2 + 2(r_2^1 + r_2^2)$, we can expect that R_1 is disconnected from R_2. Specifically, if $|\mathcal{P}^1| > 5$, $|\mathcal{P}^2| > 5$ and $d(p_0^1, p_0^2) > 0.5 + r_1^1 + r_1^2 + 2(r_2^1 + r_2^2)$, we add $DC(R_1, R_2) = 1$ to the initial knowledge base. When inconsistencies are detected, these initial interpretations are weakened, e.g., $P(\textit{Cardiff Bay},\textit{Butetown}) = 1$ may be weakened to $P(\textit{Cardiff Bay},\textit{Butetown}) \geq 0.5$ to make the knowledge base consistent. Which of the initial interpretations to weaken is based on our confidence in the extracted relation (e.g., how many sources confirm it), as well as on the number of inconsistencies that would be repaired by weakening each of the interpretations.

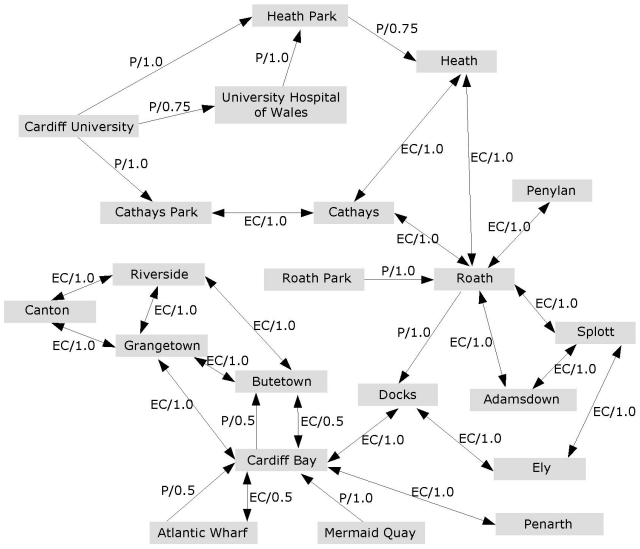

Figure 1. Fuzzy spatial relations between neighbourhoods of Cardiff.

A portion of the knowledge base that was obtained after fuzzy spatial reasoning is shown in Figure 1. For the ease of presentation, only containment and adjacency relations are displayed, and fuzzy spatial relations that follow straightforwardly from applying transitivity rules have been omitted. For example, from the fact that $P(UHW, HP) = 1$ and $P(HP, H) \geq 0.75$, we can show that also $P(UHW, H) \geq 0.75$ holds, where UHW, HP and H are abbreviations for the University Hospital of Wales, Heath Park and Heath respectively. Note that values such as 0.5 and 0.75 correspond to lower bounds for the corresponding fuzzy spatial relations. A first observation from the results in Figure 1 is that most fuzzy spatial relations are adjacency relations (EC), which could be expected since most of the place names in Table 1 are indeed non–overlapping neighbourhoods. Some notable exceptions occur, however, such as the University Hospital of Wales, Roath Park, Cardiff University, etc., but, accordingly, in most of these cases, containment relations have been found. Another observation is that most of the fuzzy spatial relations hold to degree 1. This results from the fact that only a small number of inconsistencies were

detected, which has two causes: a large fraction of the extracted relations are correct, and about some regions too little information is available to find inconsistencies. For example, $P(Roath, Docks) = 1$ is one of the few clearly wrong results. Since little information about the Docks area is available, however, this did not result in any conflicts and the error was not detected by the reasoning algorithm. Another error was introduced by the reasoning algorithm due to the ambiguity of the place name Cardiff University, part of which is located in Cathays Park. On the other hand, the University Hospital is located in Heath Park, which led to the incorrect conclusion that Heath Park and Cathays Park are overlapping (to degree 1).

Another interesting case is the relationship between Cardiff Bay and Butetown. Cardiff Bay is located towards the outskirts of an area that used to be called Tiger Bay and is more recently called Butetown, suggesting a containment relation. However, people living in or near the recently redeveloped and wealthy Cardiff Bay tend to consider their neighbourhood as disjoint from the much poorer Butetown region. Hence, both a containment relation and an adjacency relation are justified between Cardiff Bay and Butetown, to some extent. Accordingly, both fuzzy spatial relations have received a lower bound of 0.5 in the knowledge base. Similarly, both a containment relation and an adjacency relation are defensible between Atlantic Wharf and Cardiff Bay.

6. Discussion

Formally evaluating the correctness of extracted topological relations is difficult, if not impossible, as the spatial relationships between different neighbourhoods are often inherently ill–defined. The scale of our case study, being restricted to one city, should furthermore be taken into account when drawing conclusions. Nonetheless, there are a number of clear observations that can be made. First, it appears that harvesting qualitative spatial relations from the web is feasible and that a reasonable accuracy can be obtained. We cannot, however, expect the resulting description to be complete, i.e., there will always be pairs of regions whose topological relationship remains unknown. For popular neighbourhoods in the city centre, many topological relations are typically found (e.g., Roath, Cardiff Bay, ...), while this is less likely to be the case for residential neighbourhoods towards the outskirts of the city, which are often only mentioned in a very small number of web documents. Next, incorrect topological relations do not as easily lead to inconsistencies as incorrect temporal relations. For example, when only adjacency relations are found, inconsistencies can never occur. This difference with temporal information is crucial, as sufficient topological relations can only be found by heuristic techniques, relying on the subsequent reasoning step to detect errors. In the experiments described above, we partially solved this problem by adding DC relations based on available quantitative information. Thus, the chances that an error in the knowledge base leads to an inconsistency are increased, resulting in fewer unrepaired errors. Along the same lines, a significant increase in performance can be expected when topological relations would be combined with orientation and nearness information, resulting in more inferences and more detected inconsistencies.

Acknowledgment

Steven Schockaert would like to thank the Research Foundation – Flanders for funding his research (research assistant).

References

[1] E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain–text collections. In *Proceedings of ACM DL*, pages 85–94, 2000.

[2] A. Arampatzis, M. van Kreveld, I. Reinbacher, C. Jones, S. Vaid, P. Clough, H. Joho, and M. Sanderson. Web-based delineation of imprecise regions. *Computers, Environment and Urban Systems*, 30(4):436–459, 2006.

[3] C. Jones, H. Alani, and D. Tudhope. Geographic information retrieval with ontologies of place. In *Proceedings of COSIT*, pages 322–335, 2001.

[4] R. Larson. Geographic information retrieval and spatial browsing. In L. Smith and M. Gluck, editors, *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information*, pages 81–123. 1996.

[5] J. Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HTL–NAACL Workshop on Analysis of Geographic References*, pages 31–38, 2003.

[6] S. Overell, J. Magalhães, and S. Rüger. Place disambiguation with co–occurrence models. In *Working Notes of CLEF*, 2006.

[7] S. Schockaert and M. De Cock. Reasoning about vague topological information. In *Proceedings of ACM CIKM*, pages 593–602, 2007.

[8] S. Schockaert, M. De Cock, C. Cornelis, and E. Kerre. Fuzzy region connection calculus: representing vague topological information. *International Journal of Approximate Reasoning*, 48:314–331, 2008.

[9] S. Schockaert, M. De Cock, and E. Kerre. Acquiring vague temporal information from the web: towards event–based retrieval. submitted.

[10] M. Silva, B. Martins, M. Chaves, A. Afonso, and N. Cardoso. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378–399, 2006.

[11] T. Tezuka and K. Tanaka. Landmark extraction: a web mining approach. In *Proceedings of COSIT*, pages 379–396, 2005.