International Journal of Geographical Information Science
Vol. 22, No. 3, March 2008, 315–336

Research Article

Location approximation for local search services using natural language hints

S. SCHOCKAERT*, M. DE COCK and E. E. KERRE
Department of Applied Mathematics and Computer Science, Ghent University,
Krijgslaan 281, S9 9000, Gent, Belgium

(*In final form 13 August 2007*)

Local search services allow a user to search for businesses that satisfy a given geographical constraint. In contrast to traditional web search engines, current local search services rely heavily on static, structured data. Although this yields very accurate systems, it also implies a limited coverage, and limited support for using landmarks and neighborhood names in queries. To overcome these limitations, we propose to augment the structured information available to a local search service, based on the vast amount of unstructured and semi-structured data available on the web. This requires a computational framework to represent vague natural language information about the nearness of places, as well as the spatial extent of vague neighborhoods. In this paper, we propose such a framework based on fuzzy set theory, and show how natural language information can be translated into this framework. We provide experimental results that show the effectiveness of the proposed techniques, and demonstrate that local search based on natural language hints about the location of places with an unknown address, is feasible.

Keywords: Geographical information retrieval; Web intelligence; Fuzzy set theory; Local search

1. Introduction

Local search services, such as Google Maps[1], Yahoo! local[2] and MSN search local[3], allow users to search for a particular business within a certain geographic context. A user may, for example, be interested in restaurants, hotels, grocery stores, dentists, etc., which are located close to some user-specified address. Currently, such queries are evaluated against a fixed list of businesses. This allows local search services to display a high degree of accuracy, and to interact with users through a very convenient interface: query results are presented in an intuitive way, including maps that show the locations of the retrieved businesses, driving directions, user reviews, etc. Hence, it should come as no surprise that local search services have become increasingly popular.

However, local search services work in a way that is very different from traditional search engines, which use crawlers that continuously search the web for new information. The sole use of a static, structured knowledge base gives rise to a

*Corresponding author. Email: Steven.Schockaert@UGent.be

[1] http://local.google.com/

[2] http://local.yahoo.com

[3] http://search.msn.com/local/

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2008 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810701626277

number of important limitations. First of all, it restricts the coverage of the system, as many businesses will not be contained in the knowledge base, even if there are websites that contain useful information about their location. Moreover, it limits the information covered to the kind traditionally contained in the well-known yellow pages, while it is exactly the inclusion of more dynamic, ephemeral information that would bring local search services to their full potential (Himmelstein 2005). Consider, for example, a system that would also be able to deal with queries such as *give me information about concerts in Seattle during the next few weeks*, which also involves a temporal constraint.

A second limitation is related to the use of landmarks in queries. Rather than asking for *restaurants near 219 4th Ave N, Seattle*, a user might want to know about *restaurants near Space Needle*. Although the locations of important landmarks are generally available in gazetteers, many place names are not supported in current systems, e.g. because the gazetteers used do not contain any information about the place, or because the place is known by different names, not all of which are contained in the gazetteers. A related problem is the support for neighborhood names. As the boundaries of neighborhoods are usually vague, gazetteers tend to contain either no information at all about neighborhoods, or only a centroid, i.e. a location considered to be the center of the neighborhood. Consequently, local search services provide no, or very limited, support for queries such as *restaurants in Seattle's Belltown neighborhood*.

A promising solution to the aforementioned problems is to augment the available structured information with information extracted from the web. On one hand, this could be information extracted from semi-structured data. Many lists of hotels, restaurants, attractions, etc. are available on the web. Usually, it is quite easy to extract from such lists the relevant names and corresponding addresses, either by writing a wrapper manually, or by using automated wrapper induction techniques (Eikvil 1999, Kushmerick 2000). However, most information on the web is still in unstructured form, i.e. free natural language text. Because it may be very hard to find the address of a particular business or landmark in (unstructured) web pages, we can sometimes only rely on hints in natural language sentences about their location. We may know, for example, that some hotel is located in Belltown, within walking distance from Pike Place Market, and a few blocks away from Space Needle. While we cannot derive the exact location of the hotel from this, we may be able to approximate its location accurately enough to estimate the relevance w.r.t. a given query. To obtain such an approximation, we need a computational framework in which the spatial extent of a neighborhood like Belltown, and nearness information such as *within walking distance from x*, and *a few blocks away from y*, can be represented.

The aim of this paper is to show that local search using natural language hints about the location of places, is feasible. In particular, we show how nearness information in natural language, and information about the surrounding neighborhood of a place, can be translated into fuzzy restrictions, and how such fuzzy restrictions can be used to estimate the location of a place with an unknown address. In the next section, we present an overview of existing work related to the interpretation of nearness and the automatic construction of footprints. Section 3 deals with the construction of a knowledge base containing relevant geographical information, which will be used for experiments throughout the paper. Next, in Section 4, we show how nearness information in natural language can be translated into fuzzy restrictions, and how fuzzy footprints describing the (vague) spatial extent of city neighborhoods can be obtained. In Section 5, we will apply the techniques from Section 4 to estimate the location of a place. Section 6 discusses a number of experiments in which the effectiveness of the proposed techniques is demonstrated. Finally, in Section 7 some conclusions and directions for future work are presented.

2. Related work

The importance of a thorough understanding of the meaning of nearness has long been realized. Early work has mainly focused on cognitive aspects of nearness (Lundberg and Eckman 1973, Sadalla *et al.* 1980), showing, among others, that nearness is context-dependent and that cognitive distortions can occur because of the existence of landmarks. More recently, several computational models for nearness have been suggested, to some extent based on results from cognitive geography. For example, Worboys (2001) discusses three possible approaches to represent nearness: a three-valued approach, a four-valued approach, and a fuzzy approach. In the three-valued approach, the nearness of two places can either be true, false, or undecided. By analyzing the results of a questionnaire, the authors conclude that nearness is neither symmetric nor transitive, although some weakened asymmetry and transitivity properties seem to hold. An analysis based on a four-valued logic aims at finding out whether situations in which the nearness of two places is undecided result from too much information (e.g. *a* is both near and not near to *b*; truth glut hypothesis) or too little (e.g. *a* is neither near *b* nor not near *b*; truth gap hypothesis). The results from the questionnaire provide some evidence towards the truth gap hypothesis. Finally, the fuzzy approach allows differentiating between degrees of nearness. The degree of nearness of two places is based on the percentage of the participants that considered these places to be near.

Most other computational models of nearness are based on a fuzzy approach as well, but use a membership function that defines how the (e.g. Euclidean) distance of two places is related to the degree of nearness (Dutta 1991, Gahegan 1995, Guesgen and Albrecht 2000). Usually, this membership function is given as such, providing very little justification. For example, the degree of nearness of two places is either defined as the reciprocal of their Euclidean distance or assumed to be known in advance, i.e. a complete enumeration of the nearness of every pair of places is specified (Guesgen and Albrecht 2000). Also scale factors are taken into account (Gahegan 1995). Other context dependencies are discussed, but not implemented into the model; in particular, the attractiveness of objects (e.g. 1 km from a shop may be far, but 1 km from a toxic waste dump very near) and reachability. The task of finding the fuzzy restrictions on the possible positions of a set of objects, induced by an initial set of fuzzy restrictions, is discussed (Dutta 1991). The reasoning scheme proposed is based on the compositional rule of inference, a well-known technique from fuzzy logic. Robinson (2000) deals with the construction of fuzzy sets for concepts such as near and far, by asking a user a series of questions of the form *Do you consider x to be far from y*, which have to be answered by either yes or no (*x* and *y* are cities, and users are given a map to answer the questions). The goal is to allow for flexible querying in geographical information systems (GIS), by using membership definitions of vague nearness relations that correspond to the interpretation of these concepts by the user.

As most authors have been focusing on GIS systems, which are based on structured information, existing work generally deals with the concept of nearness as such. To our knowledge, the automatic construction of a computational representation for natural language nearness relations such as *within walking distance*, has not yet been addressed. However, Yao and Thill (2005) addresses the inverse problem of predicting which natural language nearness relation is most appropriate (e.g. very near, near, normal, etc.), given the exact distance and context variables. The context variables allow the statistical model to deal with factors such as scale, the type of activity, etc. While the results seem promising, the proposed technique can only be applied to this inverse problem, and not to find the (fuzzy) range of possible distances, given a natural language nearness relation.

Despite the wide interest in the concept of nearness, its application in geographical information retrieval has, until now, been rather limited. Techniques to extract the names of cognitively significant landmarks from the web are introduced in Tezuka and Tanaka (2005). One advantage of the suggested techniques is that they allow determining the significance of a landmark in a quantitative way. Alternatively, a data mining technique for finding significant place names is proposed (Duckham and Worboys 2001), also with the aim of bridging the gap between computational and cognitive approaches to nearness. It is proposed to limit the range of places near a certain landmark, based on the popularity of the landmark (Tezuka *et al.* 2001). In particular, the paper claims that the more popular a particular landmark is, the larger the area considered to be near that landmark will be. The use of nearness relations in natural language to improve geographical information retrieval was also addressed in Delboni *et al.* (2007). The aim of their work is to improve the geographical awareness of traditional search engines, by using information about landmarks and nearness relations for query expansion. The working hypothesis is that nearness relations such as near, close, in front of, etc., all have a similar meaning. Hence, a user interested in *hotels near Space Needle* is also interested in *hotels close to Space Needle*. Using this technique, they show a significant improvement in terms of precision and recall of geographically relevant web pages, compared to traditional search engines (Google was used in their experiments). The main advantages of their approach are that no gazetteers are needed, and that the proposed query expansion strategy can be applied to any traditional keyword-based search engine with minimal effort.

Another line of research relevant to our work is the automated construction of fuzzy footprints, i.e. representations of the spatial extent of vague regions or neighborhoods. The need to deal with fuzzy footprints when representing certain regions has been pointed by various authors (Goodchild *et al.* 1998, Hill *et al.* 1999, Harada and Sadahiro, 2005, Schockaert *et al.* 2005). Although the focus is generally on large-scale regions such as Western Europe, or the Alpes, the same considerations apply to city neighborhoods. An experiment is discussed in which users were asked to show on a map what they understood as downtown Santa Barbara (Montello *et al.* 2003). Based on the results of this experiment, the authors suggest that it would be feasible to construct fuzzy footprints through interaction with the users of a system. Several automatic methods to construct representations of vague regions have already been proposed. For example, techniques to find a crisp boundary for vague regions were proposed by Reinbacher *et al.* (2005), while Harada and Sadahiro (2005) presented a statistical solution to the problem. We proposed an approach based on fuzzy set theory (Schockaert *et al.* 2005), using natural language constraints found in web documents such as *x is located in the north of R*.

3. Obtaining data

As our focus is not on the extraction of spatial knowledge from the web, we have primarily used data extracted from semi-structured documents to construct a knowledge base of spatial information. One example is Hotel-Rates.com[4], which contains a list of hotels for most reasonably large cities in the world. We extracted the information in these lists by manually defining rules based on the structure of the corresponding html-documents, a technique known as screen scraping. Although this technique is very useful for the kind of experiments described in this paper, more advanced techniques would be required to implement a fully fledged local search service. One possibility is to use automated wrapper induction techniques, which try to discover the rules that would be used for screen scraping automatically [*see* (Eikvil 1999) for an overview].

Furthermore, for each hotel in the lists, a pointer to a document about the hotel is provided. These documents contain a natural language description of the hotel, as well as semi-structured information about the surrounding neighborhood and nearby attractions. To analyze the natural language description, we first parsed all relevant sentences using the Stanford Parser[5]. Then we extracted spatial relations using patterns such as

located within walking distance of <NP>
located in the heart of <NP>

For example, in a sentence such as *the hotel is located within walking distance of the University of Washington campus, and …*, the parser would correctly identify the *University of Washington campus* as a noun phrase (NP). Because this sentence therefore matches the pattern, we assume that the nearness relation *within walking distance* holds between the hotel described on the web page, and the University of Washington campus. We used a large set of patterns, covering 20 named nearness relations (e.g. across the street, in the heart of, etc.), as well as phrases expressing a number of kilometers, miles, blocks, meters, and yards. From the semi-structured information, additional nearness relations are extracted, as well as the surrounding neighborhood (when available). On average, this process resulted in 11.27 natural language hints per hotel. In a similar way, we have extracted spatial information from channels.nl[6] and from openlist[7]. From openlist, we also extracted lists of restaurants and lists of touristic attractions, as well as some useful nearness relations available in semi-structured form. In particular, for most hotels, a list of nearby restaurants and attractions is provided, as well as a number of alternative (close) hotels that could be considered. Furthermore, openlist also contains lists of places located in a particular neighborhood of the city. We used these lists to add information about the surrounding neighborhood of places to our knowledge base.

In total we extracted information about 56 US cities. The process outlined above gave us a list of over 60,000 place names (7,819 hotels, 47,152 restaurants, and 8,504 touristic attractions) with corresponding addresses, as well as spatial relations between some of the hotels and some of the attractions and restaurants. We used the

[4] http://www.hotel-rates.com/
[5] http://nlp.stanford.edu/downloads/lex-parser.shtml
[6] http://www.channels.nl/
[7] http://www.openlist.com/

geocoding service of the Google Maps API[8] to translate the addresses to geographical coordinates.

To represent the spatial extent of neighborhoods, we need a list of places assumed to lie in each neighborhood of interest. Apart from the information that is already in our knowledge base, we use information coming from two sources for this: Yahoo! local, and restaurants.com[9]. To extract relevant places from Yahoo! local, we submit a query with the name of the neighborhood as a keyword, and the name of the city as the geographical restriction. From the list of places returned, we keep the places whose name contains the name of the neighborhood (e.g. Belltown pizza is probably located in Belltown), as well as places from whose description we can find out that they are located in the neighborhood, using a pattern-based approach. The information on restaurants.com is semi-structured; we use screen scraping to extract the names and addresses of restaurants located in a particular neighborhood, as well as a list of neighborhood names for the city under consideration.

4. Representing vague geographical information

4.1 *Fuzzy restrictions*

A lot of useful geographical information in natural language takes the form of vague assertions about the nearness of two places. A question which naturally arises from this is: what can we say about the possible locations of an unknown place x, knowing only the location of a and the fact that a is, e.g. at walking distance from x? Our knowledge about the location of x is clearly vague, i.e. there exists a set of locations that are definitely compatible with this knowledge, there exists a set of locations that are definitely not compatible, and there exists a third set, consisting of borderline cases, which are neither fully compatible, nor fully incompatible.

We will use fuzzy sets and fuzzy relations to represent the vague geographical knowledge at our disposal. A fuzzy set in a universe X is formally defined as a mapping A from X to the unit interval [0, 1] (Zadeh 1965). For x in X, $A(x)$ is called the membership degree of x in A, and reflects the extent to which x has the (vague) property that A is modeling. A fuzzy set in the universe $X \times Y$ is also called a fuzzy relation from X to Y; a fuzzy relation from X to X is also called a fuzzy relation in X. Fuzzy relations are particularly useful to represent nearness relations.

Let α, β, γ, δ be non-negative real numbers such that $\alpha \leqslant \beta \leqslant \gamma \leqslant \delta$. The fuzzy relation $R_{(\alpha,\beta,\gamma,\delta)}$ in the universe of locations is defined for locations x and y as

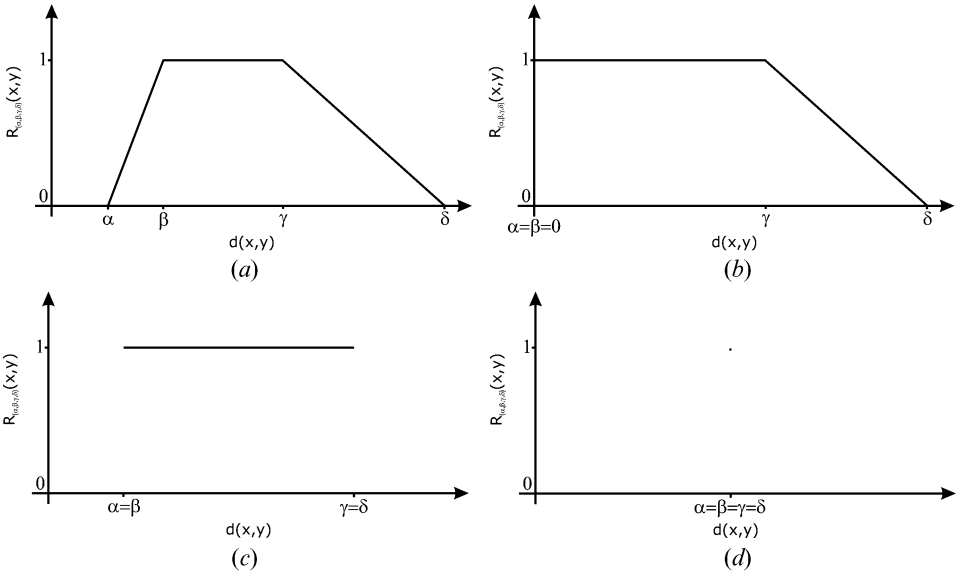
$$R_{(\alpha,\beta,\gamma,\delta)}(x,y) = \begin{cases} \dfrac{d(x,y)-\alpha}{\beta-\alpha} & \text{if } \alpha < d(x,y) < \beta \\ 1 & \text{if } \beta \leq d(x,y) \leq \gamma \\ \dfrac{\delta-d(x,y)}{\delta-\gamma} & \text{if } \gamma < d(x,y) < \delta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

[8] http://www.google.com/apis/maps/

[9] http://www.restaurants.com

where d is the straight-line distance[10]. By representing a natural language nearness relation such as *within walking distance* as a fuzzy relation $R_{(\alpha,\beta,\gamma,\delta)}$, we specify a fuzzy lower bound and fuzzy upper bound on the possible distances between places said to be within walking distance from each other. This is illustrated in figure 1a. If only an upper bound is required, we can choose $\alpha=\beta=0$, as shown in figure 1b. Figure 1c and d illustrate that also crisp restrictions, such as *between 2 and 4 km* and *exactly 1.5 km*, can be represented within this framework.

The use of trapezoidally shaped fuzzy sets to define nearness relations, in the context of this paper, offers many advantages. First of all, processing trapezoidally shaped fuzzy sets is computationally much more efficient than processing arbitrary fuzzy sets or relations. For example, sometimes we know that a place a is close to a place b, which is within walking distance from a place c. To derive useful information about the nearness of a and c from this, we need to compose a fuzzy relation representing *close to* with a fuzzy relation representing *within walking distance*. If trapezoidally shaped fuzzy sets are used, efficient characterizations of such compositions can be used (Schockaert *et al.* 2006). Furthermore, trapezoidally shaped fuzzy sets are defined using only four parameters, which have an intuitive meaning. Finally, the use of fuzzy sets with a relatively simple shape is important for the robustness of the approach. Since we use the web to obtain input data, we usually have a large amount of data available to construct an appropriate representation of a particular nearness relation. However, using the web also implies that individual samples of our input data may not be very reliable. Using trapezoidally shaped fuzzy sets allows us to sufficiently abstract away from individual input samples.



(*a*) (*b*) (*c*) (*d*)

Figure 1. Nearness relation is represented as a fuzzy restriction on the distance between the two places x and y it applies to. (a) Fuzzy distance restriction; (b) Fuzzy upper bound; (c) Crisp distance restriction; (d) Exact distance.

[10] One can think of this straight-line distance as the Euclidean distance. However, in practice usually the circle distance (i.e., the length of the shortest path between two points on the surface of a sphere) would be used instead, since locations are typically expressed as longitude and latitude coordinates.

This stands in contrast to approaches (such as Robinson 2000), in which interpretations of nearness relations are constructed by directly asking questions to human users. In such approaches, relatively little data are usually available, which is, however, very reliable. Therefore, it may be useful to use fuzzy sets with a more complex shape, which fit the actual input data more accurately, and to use prior knowledge about the human users to make decisions in the case of inconsistencies between different users.

4.2 *Representing named nearness relations*

A lot of information about the nearness of places is expressed in texts using named natural language relations such as *within walking distance*. To represent such information within the framework outlined above, we need to find appropriate values of the parameters α, β, γ, and δ, for each frequently occurring named nearness relation. To find these values, we start with a set $S=\{(p_1, q_1), (p_2, q_2), \ldots, (p_n, q_n)\}$ of pairs of places said to be within walking distance of each other, and for which we know the exact distance. In particular, let d_i be the (straight-line) distance between p_i and q_i. Without loss of generality, we can assume that $d_1 \leqslant d_2 \leqslant \ldots \leqslant d_n$. Figure 2 shows how often the distance between places from our knowledge base said to be within walking distance of each other is between 0 and 1 km, between 1 and 2 km, etc. As can be seen from this figure, the set S contains outliers, e.g. places that are more than 10 km away from each other, but are still said to be within walking distance. This can, for example, be due to errors in the phase of extracting information from web pages, the use of ambiguous place names, or incorrect geocoding of the corresponding addresses. To define the interpretation of *within walking distance*, we have to specify an interval $[\beta, \gamma]$ of distances that are fully compatible, as well as values for β–α and δ–γ which specify the degree of vagueness of the lower and upper bound, i.e. how flexible these bounds should be.

Because of the existence of outliers, we cannot choose $[\beta, \gamma]=[d_1, d_n]$. Rather, we choose 4 representative distances $d_{n_1}, d_{n_2}, d_{n_3}, d_{n_4}$, where $1 \leqslant n_1 < n_2 < n_3 < n_4 \leqslant n$ (with $n \geqslant 4$). The idea is that the distances in $\{d_1, d_2, \ldots, d_{n_1-1}\}$ and in $\{d_{n_4+1}, d_{n_4+2}, \ldots, d_n\}$ might be outliers. Furthermore, we assume that $d_{n_2} - d_{n_1}$

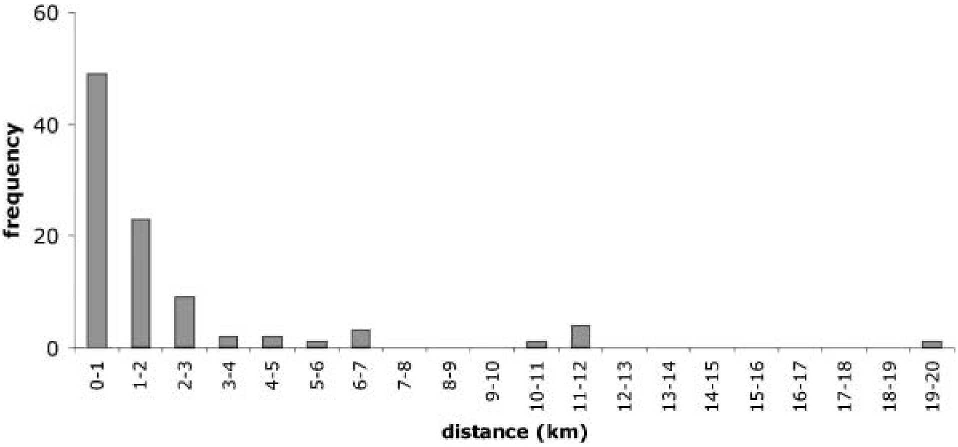


Figure 2. Frequency of distances between places said to be within walking distance of each other.

(resp. $d_{n_4} - d_{n_3}$) gives a good indication of the vagueness of the lower (resp. upper) bound. We define the parameters α, β, γ, and δ, i.e. the interpretation of *within walking distance* as

$$\alpha = d_{n_1} - a_1(d_{n_2} - d_{n_1}) \tag{2}$$

$$\beta = d_{n_1} - a_2(d_{n_2} - d_{n_1}) \tag{3}$$

$$\gamma = d_{n_4} + a_3(d_{n_4} - d_{n_3}) \tag{4}$$

$$\delta = d_{n_4} + a_4(d_{n_4} - d_{n_3}) \tag{5}$$

where $a_1 \geqslant a_2 \geqslant 0$ and $a_4 \geqslant a_3 \geqslant 0$. Large values of the parameters a_i correspond to a tolerant interpretation of the nearness relation, while small values of a_i correspond to a strict interpretation. For example, choosing $a_2 = a_3 = 0$ means that only the distances in $[d_{n_1}, d_{n_4}]$ are considered to be fully compatible with the nearness relation under consideration. Such an interpretation would probably be too strict for many applications. On the other hand, if these parameter values would be chosen too large, the resulting fuzzy relations would be too tolerant, and would therefore convey too little information. Optimal values of a_1, a_2, a_3, and a_4 depend on the kind of data used. We used $a_2 = a_3 = 1$ and $a_1 = a_4 = 3$, as initial experiments revealed that these values provided an appropriate trade-off between flexibility and informativity for the kind of data discussed in this paper. Also the optimal value of the parameters n_1, n_2, n_3, n_4 might depend on the kind of data that is used; we used $n_1 = \frac{n}{5}$, $n_2 = \frac{2n}{5}$, $n_3 = \frac{3n}{5}$ and $n_4 = \frac{4n}{5}$ (assuming n is a multiple of 5, for simplicity). This choice of parameters leads to the interpretations in table 1. For example, knowing that x is within walking distance of y, all distances between 54 m and 2.55 km are equally possible candidates for the straight-line distance. Moreover, all distances between 0 and 4.094 km are all possible to some extent. Note that when $\alpha \leqslant 0$ and $\beta \leqslant 0$, no lower bound on the possible distances is imposed. As could be expected, nearness relations such as *near* and *close* convey less information than *within walking distance* or *across the street*. However, the upper bound of *adjacent* is somewhat surprising, as one could expect that the meaning of *adjacent* would be quite similar to the meaning of *across the street*. A closer look at the data reveals that adjacent is often used w.r.t. places whose spatial extent is not negligible (e.g. parks or famous streets). However, like in most gazetteers, we have represented the location of, for example, a park, as a point, and used the distance to this point rather than to the boundary of the park. Solutions to this problem are far from obvious, since the boundaries of parks are usually not available, and automated methods to extract footprints from the web are not suitable for places such as parks.

Nearness relations cannot only be found in texts, we can also extract information about nearness from semi-structured information sources. In particular, from openlist we extracted for each hotel a list of nearby attractions and restaurants, and a list of alternative hotels that could be considered. We treated this information in the same way as natural language nearness relations; the results are also shown in table 1. Although these relations are clearly much more general than, for example, *within walking distance*, they can still be very useful, as we have a very high number of such relations at our disposal.

In the previous discussion, we have neglected the fact that the meaning of nearness relations can be dependent on the context in which they are used. Mostly this is justified because all relations actually occur in more or less the same context.

Table 1. Interpretations for some frequently occurring named nearness relations (distances in km)

Nearness relation	Frequency	d_{n_1}	d_{n_2}	d_{n_3}	d_{n_4}	α	β	γ	δ
Within walking distance	114	0.380	0.706	1.005	1.777	−0.598	0.054	2.550	4.094
Across the street	36	0.092	0.215	0.296	0.540	−0.277	−0.031	0.784	1.273
Near	39	0.456	1.205	2.077	9.210	−1.789	−0.292	16.342	30.607
Close	24	0.871	2.445	3.578	9.546	−3.849	−0.702	15.513	27.448
Adjacent	52	0.183	0.325	0.796	3.687	−0.244	0.040	6.579	12.361
Nearby (openlist.com)	12,419	0.966	1.660	2.441	3.369	−1.115	0.272	4.297	6.153
Alternates (openlist.com)	4151	1.133	2.735	5.745	11.595	−3.674	−0.469	17.444	29.143

For example, scale factors should not be taken into account because the scale is always similar, i.e. that of a large US city. Another issue is the asymmetry of nearness relations. For example, if we would extract a list of nearby hotels from the web page of a famous touristic attraction, we would have to interpret this in a different way than if we would extract a list of nearby attractions from the web page of a hotel. Again, this is not a problem when using the relations from our knowledge base, since they always express nearness from the point of view of the hotel. One factor that may be relevant, however, is the influence of the popularity of touristic attractions. As pointed out by Tezuka and Tanaka (2005), the interpretation of *near a famous place* may be less specific than *near a rather unknown place*, because, for example, hotel owners want to suggest that their hotel is close to famous places. To assess whether this claim holds for the kind of information in our knowledge base, we refined the interpretations from table 1 to those in table 2. The idea is that we calculate two sets of parameters for each nearness relation: one using only popular places, and the other using only unpopular places, where a place is defined as popular if it occurs at least 5 times as the object of a nearness relation in our knowledge base. Table 2 clearly shows that *within walking distance*, *across the street*, and the alternatives given by openlist.com, are in accordance with this claim by Tezuka and Tanaka (2005). However, the other relations display the exact opposite behavior, i.e. the interpretation of nearness seems to be narrower for popular places. One possible explanation for this could be that famous places tend to be in the city center, where hotels, restaurants, and touristic attractions are more close to each other than in the outskirts. In some experiments, we will use these refined interpretations, except for *close*, because there are too few occurrences of this relation in our knowledge base to find reliable parameters.

4.3 *Representing quantified nearness relations*

While there is already an abundance of nearness information on web pages that uses named relations, there may be even more information that expresses nearness in terms of a specific number of kilometers, miles, blocks, etc. Although a statement such as *the hotel is located at 3 kilometers from Space Needle* might seem to convey an exact distance at first glance, the intended distance restriction is vague. First of all, *at 3 km* should probably be understood as *at approximately 3 km*, since overspecific information, such as *at 3.124 km*, is generally avoided in texts. Next, it may happen that the writer of this information does not know the exact distance, and simply writes 3 km as an approximation of the real distance. Finally, it is not clear whether the 3 km restriction applies to the straight-line distance, or to the actual traveling distance. This is further complicated by the fact that we have no information about the actual traveling distance. Even using a route planner would not solve all problems, since, for example, the walking distance may differ from the traveling distance by car (e.g. due to one way streets).

Rather, we will rely on the assumption that the actual traveling distance differs from the straight-line distance by at most a factor $\sqrt{2}$. To justify this, consider a city block street layout as in figure 3a. The length of the shortest path from place a to place b is $\sqrt{2}d(a, b)$ km, where $d(a, b)$ is the straight-line distance in kilometers. The situation in figure 3a reflects the worst possible street layout (i.e. the street layout that results in the longest distance) which still has the property that there is a path for which the distance to b is decreased in every step. Especially when the straight-line distance between a and b is very small, a situation like in figure 3b can occur,

Table 2. Refined interpretations for some frequently occurring named nearness relations (distances in km). Popular places are defined as places that occur at least 5 times as the object of a nearness relation in our knowledge base.

Nearness relation		Frequency	d_{n_1}	d_{n_2}	d_{n_3}	d_{n_4}	α_0	β_0	γ_0	δ_0
Within walking distance	not popular	27	0.271	0.350	0.732	1.439	0.034	0.192	2.145	3.558
	popular	87	0.507	0.816	1.025	1.777	−0.421	0.198	2.529	4.034
Across the street	not popular	11	0.084	0.151	0.288	0.361	−0.119	0.016	0.433	0.578
	popular	25	0.176	0.225	0.303	0.569	0.030	0.128	0.835	1.366
Near	not popular	15	0.731	1.694	2.325	12.276	−2.155	−0.230	22.226	42.128
	popular	24	0.375	0.877	2.004	7.867	−1.132	−0.127	13.730	25.457
Close	not popular	4	0	0.175	3.579	11.202	−0.175	−0.525	18.825	34.071
	popular	20	1.419	3.317	3.950	9.546	−4.275	−0.479	15.142	26.334
Adjacent	not popular	20	0.103	0.442	0.796	9.396	−0.913	−0.235	17.996	35.196
	popular	32	0.189	0.282	0.989	3.687	−0.089	0.097	6.385	11.781
Nearby (openlist.com)	not popular	554	1.885	2.850	3.714	4.627	−1.010	0.920	5.541	7.369
	popular	11865	0.942	1.618	2.389	3.271	−1.083	0.268	4.154	5.920
Alternates (openlist.com)	not popular	481	0.543	1.342	3.190	6.760	−1.854	−0.256	10.330	17.481
	popular	3670	1.252	2.973	6.287	12.248	−3.912	−0.470	18.209	30.131

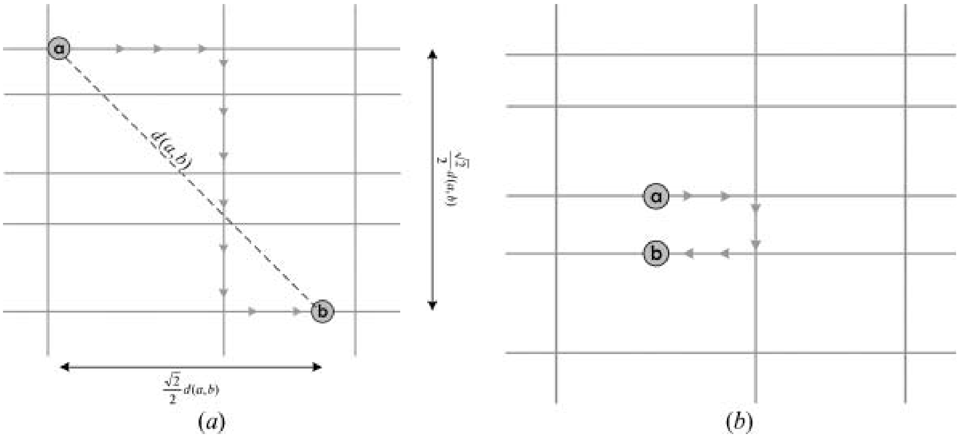


Figure 3. We assume that the actual travelling distance from a to b differs from the straight-line distance $d(a, b)$ by at most a factor $\sqrt{2}$.

where all paths to b pass at some point c where the straight-line distance to b is greater than from a. To cope with this, we will treat small distances in a different way, as is explained below.

To find appropriate values of the parameters α, β, γ, and δ for a nearness relation such as *3 km from* we assume that $\alpha=3\alpha_0$, $\beta=3\beta_0$, $\gamma=3\gamma_0$, $\delta=3\delta_0$, where the parameters α_0, β_0, γ_0, and δ_0 are the same for all nearness relations of the form *r km from* ($r \in]0, +\infty[$). To determine the values of α_0, β_0, γ_0, and δ_0, we proceed as in Section 4.2, using the distances $d_1, d_2, \ldots, d_n$ obtained by dividing the straight-line distance of every two places said to be at r km from each other, by r. In other words, rather than modeling *r km from*, we model *1 km from*, and multiply the parameters obtained by r. The resulting parameters, modeling in fact *1 km from*, *1 mile from*, and *1 block from*, are shown in table 3. Note that the ranges of possible distances entail those that could be expected from the argument above, i.e. $\left[\frac{1}{\sqrt{2}}, 1\right] = [0.707, 1]$ for kilometer, and $\left[\frac{1.6093}{\sqrt{2}}, 1.6093\right] = [1.138, 1.6093]$ for mile. Also, the ranges are quite vague, resulting from the fact that the distances mentioned in texts are often approximations and the fact that, for example, hotel owners are not always fully honest about the true location of their hotel. Furthermore, note that the range of possible distances for 1 block from is vaguer than those for *1 km from* or *1 mile from*. This is due to the fact that a block is an inherently vague unit, unlike a kilometer or a mile.

Another reason for the deviation from the ranges [0.707, 1] and [1.138, 1.6093] are the simplifications we made w.r.t. reachability, i.e. the difference between the straight-line distance and the actual traveling distance. As argued above, we can

Table 3. Interpretations for some frequently occurring quantified nearness relations (distances in km, $r \in]0, +\infty[$, $k \in \mathcal{N}\backslash\{0\}$).

Nearness relation	Frequency	d_{n_1}	d_{n_2}	d_{n_3}	d_{n_4}	α_0	β_0	γ_0	δ_0
r kilometer(s)	785	0.647	0.816	0.967	1.202	0.138	0.477	1.437	1.908
r mile(s)	3,063	1.003	1.248	1.478	1.806	0.270	0.759	2.135	2.792
k block(s)	672	0.102	0.142	0.224	0.811	−0.020	0.061	1.397	2.571

expect this to be particularly true for small distances. Our proposed solution is to use different sets of parameters for small distances. The results of this are shown in table 4, which confirm our idea that small distances behave in a different way. For example, all distances in the range [0.15, 1.45] are fully compatible with the nearness relation *0.1 km from*, while the distances in [0.70, 2.79] and [4.84, 11.67] are fully compatible with the nearness relations *1 km from* and *10 km from* respectively.

4.4 *Representing neighborhoods*

Semi-structured and unstructured information usually contains a lot of information about the neighborhood in which a particular place is located. Like information about the nearness to other places, information about the surrounding neighborhood of a place could be very useful to find an approximation of its location. However, this requires a representation of the spatial extent, i.e. a footprint, of city neighborhoods. As the boundaries of such neighborhoods are typically vague, gazetteers either contain no information at all about neighborhoods, or provide only a centroid (i.e. the coordinates of a single place considered to be the center of the neighborhood). To be able to use neighborhood information, we will therefore try to find information about the boundaries of neighborhoods automatically.

Recall from Section 3 that our knowledge base contains, for each neighborhood of interest, a set $L=\{l_1, l_2, \ldots, l_m\}$ of places assumed to lie in the neighborhood. We will use this information to construct a fuzzy footprint, i.e. a fuzzy set F in the universe of locations, such that $F(x)$ expresses the degree to which a location x is contained in the neighborhood. Let l^* be the medoid of the set L, i.e. the place of L for which the sum of the distances to the other places is minimal

$$l^* = \operatorname*{argmin}_{l_1 \in L} \sum_{l_2 \in L \setminus \{l_1\}} d(l_1, l_2) \tag{6}$$

Without loss of generality, we can assume that $d(l^*, l_1) \leqslant d(l^*, l_2) \leqslant \ldots \leqslant d(l^*, l_m)$. Our main idea to find a fuzzy footprint is very similar to the way we constructed the interpretations for the nearness relations. In particular, we assume that at most 40% of the locations in L are noisy (i.e. incorrectly classified as lying in the neighborhood), and that the difference $d\left(l^*, l_{\frac{3m}{5}}\right) - d\left(l^*, l_{\frac{2m}{5}}\right)$ gives a good indication of the vagueness of the boundaries of the neighborhood (where we assume that m is a multiple of 5, for simplicity). For locations l in L, we define F as

$$F(l) = \begin{cases} 1 & \text{if } d(l^*, l) \leq \lambda \\ 0 & \text{if } d(l^*, l) \geq \rho \\ \dfrac{\rho - d(l^*, l)}{\rho - \lambda} & \text{otherwise} \end{cases} \tag{7}$$

where

$$\lambda = d\left(l^*, l_{\frac{3m}{5}}\right) \tag{8}$$

$$\rho = d\left(l^*, l_{\frac{3m}{5}}\right) + 4\left(d\left(l^*, l_{\frac{3m}{5}}\right) - d\left(l^*, l_{\frac{2m}{5}}\right)\right) \tag{9}$$

Note that (at least) 60% of the locations in L is assumed to lie in the neighborhood to degree 1. The definition of F for locations l that are not contained in L is based on

Table 4. Refined interpretations for some frequently occurring quantified nearness relations (distances in km).

Nearness relation		Frequency	d_{n_1}	d_{n_2}	d_{n_3}	d_{n_4}	α_0	β_0	γ_0	δ_0
r kilometer(s)	$r \in]0, 0.5]$	26	2.133	2.740	4.730	9.614	0.314	1.527	14.498	24.268
	$r \in]0.5, 1]$	54	0.888	1.075	1.390	2.089	0.326	0.701	2.789	4.189
	$r \in]1, 5]$	269	0.619	0.784	0.987	1.220	0.124	0.454	1.453	1.919
	$r \in]5, +\infty[$	436	0.637	0.790	0.920	1.043	0.177	0.484	1.167	1.412
r mile(s)	$r \in]0, 0.5]$	260	1.470	1.864	2.862	6.463	0.288	1.076	10.063	17.265
	$r \in]0.5, 1]$	476	1.076	1.378	1.647	2.091	0.169	0.773	2.535	3.423
	$r \in]1, 5]$	1030	1.006	1.213	1.470	1.782	0.384	0.798	2.094	2.717
	$r \in]5, +\infty[$	1297	0.896	1.177	1.361	1.560	0.054	0.615	1.760	2.158
k block(s)	$k=1$	97	0.171	0.299	1.019	5.100	−0.212	0.043	9.181	17.343
	$k=2$	113	0.128	0.210	0.306	1.745	−0.118	0.046	3.184	6.063
	$k=3$	95	0.112	0.153	0.230	1.532	−0.010	0.071	2.836	5.441
	$k>3$	367	0.092	0.117	0.164	0.328	0.015	0.066	0.492	0.820

the convex hull of particular subsets of L. For $k \in]0, 1]$, we define the set M_k of locations as the convex hull of the locations l of L for which $F(l) \geqslant k$. Finally, for an arbitrary location l (i.e. l not necessarily in L), we define F as

$$F(l) = \sup\{k | k \in]0, 1] \text{ and } l \in M_k\} \tag{10}$$

For example, assume that $L = \{l_1, l_2, l_3, l_4, l_5, l_6\}$, $l_1 = l^*$, and that $F(l_1) = F(l_2) = F(l_3) = 1$, $F(l_4) = 0.9$, $F(l_5) = 0.8$, and $F(l_6) = 0.7$. The resulting definitions of the sets M_1, $M_{0.8}$, and $M_{0.7}$ are shown in figure4. For any location l, $F(l) = 1$ provided $l \in M_1$, while $F(l) = 0.9$ iff $l \in M_{0.9} \backslash M_1$, $F(l) = 0.8$ iff $l \in M_{0.8} \backslash M_{0.9}$, $F(l) = 0.7$ iff $l \in M_{0.7} \backslash M_{0.8}$, and $F(l) = 0$ otherwise.

Schockaert *et al.* (2005) propose a number of techniques to refine the fuzzy footprint of a large-scale region R are proposed, based on natural language information such as *x is located in the north of R*. We did not apply these techniques here because this kind of natural language information is less abundant for neighborhoods than it is for large-scale regions, and because the number of places available for each neighborhood (i.e. the cardinality of L) is sufficiently high to allow for simpler techniques.

5. Location approximation

In Section 4, we explained how natural language hints such as *x is located within walking distance from a*, and *x is located in N* could be interpreted. If a is a place with a known location, and N a neighborhood with a known fuzzy footprint, these hints can be translated to fuzzy sets, defining which places are compatible with them, and to what extent. In this section, we will show how the location of x can be estimated, using only natural language hints that relate x to places with a known location or fuzzy footprint.

Because we cannot assume that all information about x is consistent, we will first identify the locations that are consistent with as much of our information as possible. Let $A_1, A_2, \ldots, A_n$ be the fuzzy sets of locations obtained by interpreting all natural language hints about the location of x. For information about the surrounding neighborhood such as *x is located in N*, this is the fuzzy footprint of the region N defined in Section 4.4. For nearness information such as *x is located within walking distance from a*, this is the fuzzy set A_i defined for all locations l by $A_i(l) = R_{(\alpha,\beta,\gamma,\delta)}(l, a)$, where the parameters $(\alpha, \beta, \gamma, \delta)$ are those that correspond to our interpretation of *within walking distance*. We define the score of a location l as

$$score(l) = \sum_{i=1}^{n} A_i(l) \tag{11}$$

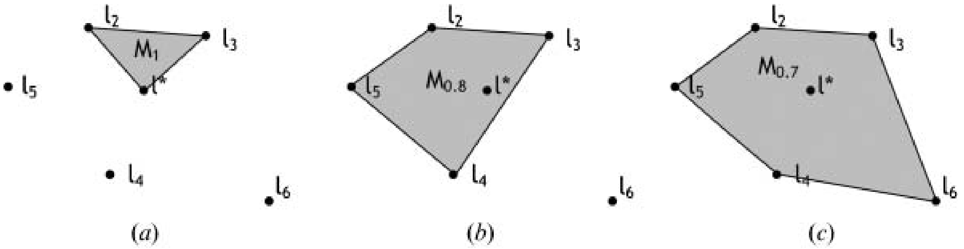


Figure 4. Definition of the sets (a) M_1, (b) $M_{0.8}$ and (c) $M_{0.7}$

This score reflects how compatible the location is with the available knowledge. Note how the use of fuzzy sets provides the flexibility needed when combining different constraints. Assume for example that there are only two fuzzy sets A_1 and A_2. If there exist locations l such that $A_1(l)=1$ and $A_2(l)=1$, we will prefer such locations. In this case our information about x is maximally consistent. When such locations do not exist, we will prefer locations that maximize $A_1(l)+A_2(l)$. Using crisp sets, we would either not have found any location that is consistent with both A_1 and A_2 in the second case (i.e. all locations in the crisp sets A_1 and A_2 would have a maximal score), or not have been able to differentiate between optimal and sub-optimal locations in the first case.

Let S be the set of locations l whose score is maximal (i.e. such that there are no locations with a higher score). This set of locations identifies a region in the real plane, which is usually not convex, and may consist of several disconnected pieces. As the estimation l_0 of the location of x, we will choose a central location from the set S. In particular, we consider a set of, e.g. 100 points that are uniformly chosen in the region identified by S. We define l_0 as the medoid of this set, as defined in equation (9).

In the following, we will use four different techniques to estimate the location of a place, three based on the procedure outlined above, and one baseline:

(i) fuzzy-1: we use the aforementioned procedure, where neither named nearness relations nor quantified nearness relations are interpreted using the refined interpretations, i.e. nearness relations are interpreted like in tables 1 and 3;

(ii) fuzzy-2: same as fuzzy-1, but the refined interpretations are used for quantified nearness relations, i.e. nearness relations are interpreted like in tables 1 and 4;

(iii) fuzzy-3: same as fuzzy-2, but the refined interpretations are also used for named nearness relations, i.e. nearness relations are interpreted like in tables 2 and 4.

As a baseline technique, we estimate the location of x without interpreting the nearness relations, and without using fuzzy footprints for neighborhoods. The idea is that every natural language hint is mapped to a single location. Information such as *x is located within walking distance of y* is mapped to the location of y, and information such as *x is located in R* is mapped to a central location of the region R. Let $Y=\{y_1, y_2, \ldots, y_n\}$ be the set of locations obtained in this way. As an estimation of the location of x, the baseline system will choose the center of gravity l_0 of Y, i.e.

$$l_0=\frac{1}{n}\sum_{\mathrm{y}\in\mathrm{Y}} y \tag{12}$$

where locations are assumed to be represented as vectors of coordinates. The purpose of using this baseline system is to evaluate how much the performance of the systems fuzzy-1, fuzzy-2, and fuzzy-3 is affected by the actual interpretation of the nearness relations and the representation of neighborhoods.

6. Experimental results

6.1 *Location approximation*

As a first evaluation of the four systems, we tried to estimate the location of hotels and touristic attractions in a number of cities, using natural language hints and the locations of the other places. In other words, to estimate the location of a hotel or an attraction, we assume that the locations of all other hotels and attractions, as well as the restaurants in our knowledge base, are known. To obtain a fair evaluation, the parameters used for the interpretation of the nearness relations were determined without using the locations in Seattle for the experiments involving Seattle locations, and similar for the other cities. Thus a different set of parameters was used for each city.

Table 5 displays the median of the straight-line distance between the estimated location of hotels and touristic attractions, and the actual location. We used the median instead of the average, because the average is too much influenced by outliers to be useful here. A first observation is that the baseline system actually performs quite well. Nonetheless, the results in table 5 clearly show that a significant improvement over the baseline was achieved by the systems fuzzy-1, fuzzy-2, and fuzzy-3, which was also confirmed by a Wilcoxon signed ranks test ($p<0.001$). This suggests that the increased complexity due to the interpretation of nearness relations, and the use of fuzzy footprints is justified for the task of location approximation. However, refining the interpretations of the nearness relations does not seem to improve the performance, i.e. the overall performance of fuzzy-1 is not worse – even slightly better – than the performance of fuzzy-2 and fuzzy-3 (Wilcoxon signed ranks, $p<0.001$).

6.2 *Local search*

In a local search setting, the system has to provide a ranking of, for example, hotels, which are near a given landmark. Ideally, the hotels in such a ranking are ordered by increasing distance from the landmark, i.e. the first hotel in the list returned by the local search service is the hotel closest to the landmark, the second hotel is the second closest hotel, etc. To assess how well our system performs at the task of finding such a ranking, we used the Spearman rank coefficient, which is well-suited to measure the correlation between different rankings of search results (Bar-Ilan 2005). The rankings we used in this experiment were obtained using the estimated locations of the hotels, and the exact locations of the touristic attractions. For each attraction *a* in each of the cities, we considered a query *hotels near a*, and calculated

Table 5. Median of the straight-line distance (in km) between the actual locations of hotels and touristic attractions, and the approximated location.

	Hotels				Attractions			
	Baseline	Fuzzy-1	Fuzzy-2	Fuzzy-3	Baseline	Fuzzy-1	Fuzzy-2	Fuzzy-3
Atlanta	3.34	1.89	1.97	1.97	4.34	1.74	1.74	1.73
Boston	1.33	1.01	0.97	0.97	1.51	0.91	1.39	1.50
Chicago	1.98	0.67	0.72	0.69	2.19	0.83	1.24	1.42
Las Vegas	2.21	1.35	1.62	1.44	2.94	1.42	1.74	1.89
Los Angeles	2.16	1.53	1.54	1.55	2.47	1.75	1.79	1.81
Miami	2.10	1.51	1.46	1.57	4.06	3.10	3.73	3.28
Minneapolis	3.84	1.55	1.53	1.52	2.02	1.98	1.98	2.51
New York	1.36	0.74	0.75	0.73	1.04	0.85	0.73	0.73
Philadelphia	1.51	1.34	1.43	1.40	2.32	1.19	1.18	1.33
Sacramento	3.44	2.34	2.07	2.50	2.59	1.23	1.35	1.73
San Francisco	1.02	0.46	0.48	0.45	1.47	0.64	0.75	0.58
Seattle	2.08	0.98	1.07	1.08	2.27	1.27	1.45	1.53

the Spearman rank coefficient between the ranking obtained with the estimated hotel locations, and the optimal ranking, i.e. the ranking obtained using the exact locations. Note that only the rankings are evaluated, and not the position at which to cut off the list of businesses returned. Therefore, the results are independent of the particular nearness relation used in the query. The results are shown in table 6. The Spearman rank coefficient is always between -1 and 1, where 1 means a perfect correlation (i.e. the rankings are identical), 0 means no correlation at all, and -1 means a perfect negative correlation. The conclusions are similar as for table 5: the behavior of fuzzy-1, fuzzy-2, and fuzzy-3 are very similar, and outperform the baseline system.

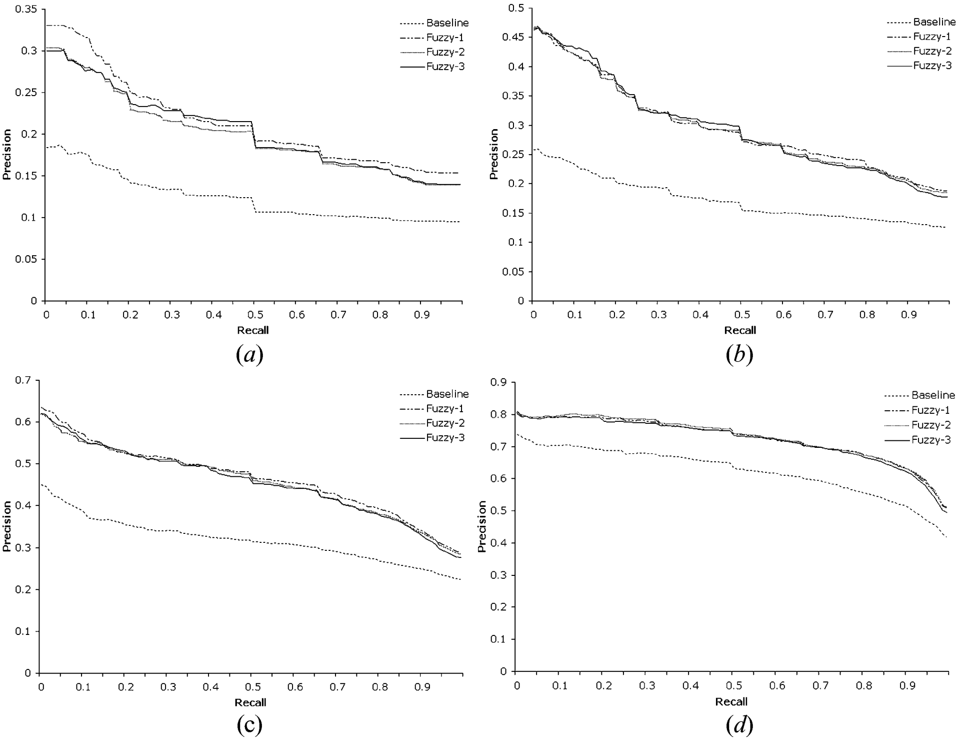
One disadvantage of using the Spearman rank coefficient is that the meaning of the results is not very intuitive: how useful is a ranking whose correlation coefficient w.r.t. the optimal ranking is 0.75? A more intuitive way of evaluating the rankings is in terms of the well-known precision and recall measures. However, this requires that we know which hotels are relevant to a query such as *hotels near a*. In the experiments, we assumed that a hotel is relevant to the query iff its location is within a fixed radius of *a*. Figure 5 shows the precision–recall curves for 4 different radiuses. The queries we considered were again *hotels near a* for each touristic attraction *a* in each of the cities. The values shown in figure 5 are averaged over all these queries. For example, when a 3 km radius is used, the precision at a recall level of 0.5 is about 0.75 for fuzzy-1, fuzzy-2, and fuzzy-3. This means that if a user would go through the list of returned hotels until she has seen half of the relevant hotels, 25% of the hotels she looked at would not have been relevant. Again, fuzzy-1, fuzzy-2, and fuzzy-3 display a similar behavior, which is significantly better than the baseline system.

7. Concluding remarks

We have discussed techniques to represent natural language nearness relations such as *within walking distance* (named nearness relations), and *3 km from* (quantified nearness relations), as well as a technique to obtain fuzzy footprints of city neighborhoods. Although these problems have, to some extent, been studied before in the context of GIS, we have focused specifically on the problem of approximating the location of a place, using data from the web. This could be very useful to

Table 6. Average Spearman rank correlation between the hotel rankings obtained using the estimated locations, and using the exact locations.

	Baseline	Fuzzy-1	Fuzzy-2	Fuzzy-3
Atlanta	0.53	0.73	0.73	0.74
Boston	0.57	0.63	0.63	0.63
Chicago	0.34	0.75	0.76	0.75
Las Vegas	0.44	0.60	0.58	0.59
Los Angeles	0.76	0.87	0.87	0.86
Miami	0.62	0.71	0.70	0.70
Minneapolis	0.50	0.73	0.73	0.70
New York	0.57	0.72	0.73	0.73
Philadelphia	0.56	0.68	0.66	0.65
Sacramento	0.51	0.69	0.67	0.66
San Francisco	0.26	0.59	0.61	0.62
Seattle	0.59	0.83	0.81	0.80



(*a*) (*b*) (c) (*d*)

Figure 5. Precision–recall curves for a query like *hotels near a*, where hotels are considered relevant to the query iff they are located at most (a) 0.25 km, (b) 0.5 km, (c) 1 km, or (d) 3 km away from *a*.

improve the performance of local search services, both in terms of increased coverage, and more flexible querying, using landmarks and neighborhood names to express desired constraints. It turns out that the nature of this task is rather different from earlier work on the interpretation of nearness relations in the context of GIS, which mainly focuses on the concept of nearness as such, and not on the actual meaning of natural language expressions to describe nearness. Also, the use of web data requires robust techniques, as the input data may be rather noisy. The model we proposed is based on trapezoidally shaped fuzzy sets, which are computationally more interesting than arbitrary fuzzy sets. Another advantage of trapezoidally shaped fuzzy sets is that they can be defined using only 4 real-valued parameters with a very intuitive meaning.

Experimental results show that local search using natural language hints, when precise locations are missing, is indeed feasible, and that the interpretations of these hints proposed in this paper are meaningful. We also looked at two techniques to refine our initial model: differentiating between small and large distances for quantified nearness relations, and between popular and unpopular places for named nearness relations. The resulting interpretations correspond closely to our intuition in the case of the quantified nearness relations. However, the presumed dependency of the named nearness relations on the popularity of the landmark involved seems to hold only for some of the named nearness relations. Moreover, the experimental results revealed no improvement, for neither of the two kinds of refinements.

To obtain a fully fledged local search service, more work on extracting spatial information from the web is needed, both to increase the coverage and to improve the accuracy of the location approximation. Furthermore, as natural language nearness relations can have different meanings in different contexts, multiple fuzzy relations may be associated with each nearness relation. Therefore, explicit representations of the contexts corresponding to each of these fuzzy relations are required, as well as heuristics to determine the context in which a particular occurrence of a nearness relation is used. Another important direction for future work is spatial reasoning. Reasoning with nearness relations could be useful to deduce extra information about the location of places. Apart from nearness relations, directional and topological information could also be used. Directional information can either refer to the relative positioning of two places, e.g. *x is located north of y*, or to the positioning of a place within a neighborhood, e.g. *x is located in the north of R*. Topological information usually refers to relations between regions, e.g. R_1 *is bordering on* R_2. Although reasoning with nearness, directional and topological information is well-studied, more work is needed on combining this in one framework.

References

Bar-Ilan, J., 2005, Comparing rankings of search results on the web. *Information Processing and Management: an International Journal*, **41**, pp. 1511–1519.

Delboni, T.M., Borges, K.A.V., Laender, A.H.F. and Davis Jr., C.A.D., 2007, Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS*, **11**, pp. 377–397.

Duckham, M. and Worboys, M., 2001, Computational structure in three-valued nearness relations. In *Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science (COSIT 2001)*, pp. 76–91 (Morro Bay, CA: Springer).

Eikvil, L., 1999, Information extraction from World Wide Web — a survey —. *Technical Report.* Available online at: http://www.nr.no/files/samba/bamg/webIE_rep945.ps (accessed 30 August 2007).

Dutta, S., 1991, Approximate spatial reasoning: integrating qualitative and quantitative constraints. *International Journal of Approximate Reasoning*, **5**, pp. 307–331.

Gahegan, M., 1995, Proximity operators for qualitative spatial reasoning. *Proceedings of the International Conference on Spatial Information Theory: a Theoretical Basis for GIS (COSIT 1995)*, pp. 31–44 (Semmering: Springer).

Goodchild, M.F., Montello, D.R., Fohl, P. and Gottsegen, J., 1998, Fuzzy spatial queries in digital spatial data libraries. In *Proceedings of the IEEE World Congress on Computational Intelligence*, vol. 1, pp. 205–210 (Anchorage: IEEE).

Guesgen, H.W. and Albrecht, J., 2000, Imprecise reasoning in geographic information systems. *Fuzzy Sets and Systems*, **113**, pp. 121–131.

Harada, Y. and Sadahiro, Y., 2005, A quantitative model of place names as a geo-referencing system. In *Proceedings of GeoComputation*, Michigan, USA. Available online at: http://igre.emich.edu/geocomputation2005/abstract_list/1200094.pdf (accessed 30 August 2007).

Hill, L.L., Frew, J. and Zheng, Q., 1999, Geographic names: the implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, **5**. Available online at: http://www.dlib.org/dlib/january99/hill/01hill.html (accessed 30 August 2007).

Himmelstein, M., 2005, Local search: the internet is the yellow pages. *Computer*, **38**, pp. 26–34.

Kushmerick, N., 2000, Wrapper induction: efficiency and expressiveness. *Artificial Intelligence*, **118**, pp. 15–68.

LUNDBERG, U. and ECKMAN, G., 1973, Subjective geographic distance: a multidimensional comparison. *Psychometrika*, **38**, pp. 113–122.

MONTELLO, D.R., GOODCHILD, M.F., GOTTSEGEN, J. and FOHL, P., 2003, Where's downtown?: behavioral methods for determining referents of vague spatial queries. *Spatial Cognition and Computation*, **3**, pp. 185–204.

REINBACHER, I., BENKERT, M., VAN KREVELD, M.J., MITCHELL, J.S.B. and WOLF, A., 2005, Delineating boundaries for imprecise regions. In *Proceedings of the 13th European Symposium on Algorithms*, pp. 143–154 (Palma de Mallorca: Springer).

ROBINSON, V.B., 2000, Individual and multipersonal fuzzy spatial relations acquired using human-machine interaction. *Fuzzy Sets and Systems*, **113**, pp. 133–145.

SADALLA, E.K., BURROUGHS, W.J. and STAPLIN, L.J., 1980, Reference points in spatial cognition. *Journal of Experimental Psychology: Human Learning and Memory*, **6**, pp. 516–528.

SCHOCKAERT, S., DE COCK, M. and KERRE, E.E., 2005, Automatic acquisition of fuzzy footprints. In *Proceedings of the Workshop on Semantic-based Geographical Information Systems*, 2006 ACM SIGIR conference, pp. 1077–1086 (Agia Napa: Springer).

SCHOCKAERT, S., DE COCK, M. and KERRE, E.E., 2006, Towards fuzzy spatial reasoning in geographic IR systems. In *Proceedings of the 3rd Workshop on Geographic Information Retrieval*, pp. 34–36 (Seattle, WA).

TEZUKA, T., LEE, R., TAKAKURA, H. and KAMBAYASHI, Y., 2001, Models for conceptual geographical prepositions based on web resources. *Journal of Geographic Information and Decision Analysis*, **5**, pp. 83–94.

TEZUKA, T. and TANAKA, K., 2005, Landmark extraction: a web mining approach. In *Proceedings of the International Conference On Spatial Information Theory (COSIT 2005)*, pp. 379–396 (New York, USA: Springer).

WORBOYS, M.F., 2001, Nearness relations in environmental space. *International Journal of Geographical Information Science*, **15**, pp. 633–651.

YAO, X. and THILL, J.-C., 2005, How far is too far? – A statistical approach to context-contingent proximity modeling. *Transactions in GIS*, **9**, pp. 157–178.

ZADEH, L.A., 1965, Fuzzy sets. *Information Sciences*, **8**, pp. 338–353.