Towards Fuzzy Spatial Reasoning in Geographic IR Systems

Steven Schockaert*, Martine De Cock, Etienne E. Kerre
Department of Applied Mathematics and Computer Science
Ghent University, Krijgslaan 281 – S9, 9000 Gent, Belgium
{Steven.Schockaert,Martine.DeCock,Etienne.Kerre}@UGent.be

ABSTRACT
Vague spatial information such as "x is located at walking distance of y" is abundant on the web. In this contribution, we propose a framework to represent such spatial information, and show how new spatial relations may be deduced. Furthermore, we illustrate how this framework can be useful to increase the coverage of focused spatial information retrieval systems.

Categories and Subject Descriptors
I.2 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

General Terms
Fuzzy spatial reasoning

1. INTRODUCTION AND MOTIVATION
There is an increasing interest in information retrieval (IR) systems that are capable of finding instances on the web of hotels, restaurants, etc. that satisfy a given spatial restriction. The need for such systems is witnessed by recent implementations by most major search engine companies: Google Maps[1], Yahoo! local[2] and MSN search local[3]. One of the main drawbacks of these existing systems is their limited coverage.

Consider a user who is interested in hotels near the University of Washington Campus in Seattle. One way to tackle this problem is to construct, offline, lists of instances of hotels and landmarks, together with their geographical coordinates, which can be obtained by geocoding their address or from an external gazetteer. Standard information extraction techniques can be used to obtain the address of a hotel or landmark from the web. Finding hotels that satisfy the user's need is then simply a matter of comparing the coordinates of the university campus with those of the hotels in the knowledge base. There are at least two scenarios in which this approach fails, resulting in a limited coverage: we may not be able to extract a correct address for some hotels or landmarks, and the geocoding of the address may go astray. Web pages of hotels, however, typically contain sentences like "The hotel is conveniently located *at walking distance* from the Space Needle". Such information gives us a rough idea of the distance between the hotel under consideration and the Space Needle. Furthermore, considering the fact that the Experience Music Project is located *near* the Space Needle, we find that the hotel must be somewhat near to the Experience Music Project as well.

In this contribution, we propose a new technique to increase the coverage of geographic IR systems, based on relevant spatial information deduced from natural language descriptions on web pages. To support reasoning of this kind, several challenges need to be addressed. The main obstacle, however, is vagueness, which may occur both at the level of the relations expressed (e.g., at walking distance from, at the doorstep of, ...) and at the level of the regions involved (e.g., downtown Seattle, the Port of Seattle, ...). We introduce a framework for representing such vague spatial knowledge, and discuss how this framework can be used in geographical IR systems. A transitivity table is introduced to support efficient reasoning.

2. FUZZY SPATIAL RELATIONS
The fuzzy footprint of a vague region is a mapping A from $\mathbb{R}^2$ to $[0,1]$ such that $A(u,v) = 1$ (at least) for some (u,v) in $\mathbb{R}^2$. For convenience, we will use the same notation A to refer both to this $\mathbb{R}^2-[0,1]$ mapping and to the region represented by it. For (u,v) in $\mathbb{R}^2$, $A(u,v) = 1$ if (u,v) definitely belongs to the region A and $A(u,v) = 0$ if (u,v) definitely not belongs to A. If the region is vague, $A(u,v) \in]0,1[$ will hold for some points (u,v), corresponding to borderline cases that belong more or less to A.

Closeness between points can be modelled as a general resemblance relation [2]. For example, for each $\alpha \in \mathbb{R}$ and $\beta \geq 0$, we define the resemblance relation $R_{(\alpha,\beta)}$ in $\mathbb{R}^2 \times \mathbb{R}^2$

*Research assistant of the Research Foundation – Flanders
[1] http://local.google.com/
[2] http://local.yahoo.com
[3] http://search.msn.com/local/

as[4]

$$R_{(\alpha,\beta)}(p,q) = \begin{cases} 1 & \text{if } d(p,q) \leq \alpha \\ 0 & \text{if } d(p,q) > \alpha+\beta \\ \frac{\alpha+\beta-d(p,q)}{\beta} & \text{otherwise } (\beta > 0) \end{cases} \quad (1)$$

for every p and q in $\mathbb{R}^2$; d is a suitable metric, e.g., the euclidean distance. Closeness between regions can be expressed through the closeness of the points that they contain. To this end we use the fuzzy logical conjunction and the fuzzy logical implication, respectively defined as $T_W(a,b) = \max(0, a + b - 1)$ and $I_W(a,b) = \min(1, 1-a+b)$ for all a and b in $[0,1]$. For example, for fuzzy footprints A and B,

$$C_R(A,B) = \sup_{p\in\mathbb{R}^2} T_W(A(p), \sup_{q\in\mathbb{R}^2} T_W(B(q), R(p,q)))$$

where $R = R_{(\alpha,\beta)}$ for some α in $\mathbb{R}$ and $\beta \geq 0$, expresses the highest degree to which there exist close points p and q, such that p belongs to A and q belongs to B. This is interpreted as the degree to which A is close to B. Furthermore,

$$P(A,B) = \inf_{p\in\mathbb{R}^2} I_W(A(p), B(p))$$

expresses the degree to which every point in A is also in B, i.e., the degree to which A is located in B. Finally,

$$NTP_R(A,B) = \inf_{p\in\mathbb{R}^2} I_W(A(p), \inf_{q\in\mathbb{R}^2} I_W(R(p,q), B(q)))$$

expresses the degree to which every point that is close to A is also close to B, i.e., the degree to which A is located in the centre of B. The correspondence of the fuzzy relations C_R, P and NTP_R with the RCC–8 calculus [3] is shown in [4]; the abbreviations used here originate from that framework.

Assume that we know that A is close to B w.r.t. $R_{(\alpha_1,\beta_1)}$, and that B is close to C w.r.t $R_{(\alpha_2,\beta_2)}$. From this, we would like to establish some knowledge about the closeness of A to C. To compose resemblance relations, we use operators that are well known in fuzzy relation calculus. In particular, for resemblance relations R and S in $\mathbb{R}^2$, $R \circ S$ and $R \lhd S$ are defined for all p and q in $\mathbb{R}^2$ by [1]:

$$(R \circ S)(p,q) = \sup_{r\in\mathbb{R}^2} T_W(R(p,r), S(r,q))$$
$$(R \lhd S)(p,q) = \inf_{r\in\mathbb{R}^2} I_W(R(p,r), S(r,q))$$

An important advantage of using resemblance relations as in (1), is that we can prove a useful characterization to efficiently evaluate $R_{(\alpha_1,\beta_1)} \circ R_{(\alpha_2,\beta_2)}$ and $R_{(\alpha_1,\beta_1)} \lhd R_{(\alpha_2,\beta_2)}$ ($\alpha_1, \alpha_2 \in \mathbb{R}$ and $\beta_1, \beta_2 \in [0,1]$). If $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$, it holds that

$$R_{(\alpha_1,\beta_1)} \circ R_{(\alpha_2,\beta_2)} = R_{(\alpha_1+\alpha_2,\max(\beta_1,\beta_2))} \quad (2)$$
$$R_{(\alpha_1,\beta_1)} \lhd R_{(\alpha_2,\beta_2)} = R_{(\alpha_2-\alpha_1+\max(0,\beta_1-\beta_2),\max(\beta_1,\beta_2))} \quad (3)$$

Similar characterizations can be shown for the case where $\alpha_1 < 0$ or $\alpha_2 < 0$; we omit the details. In the next section we explain how these compositions can be used to derive information on the closeness of A to C, based on the closeness of each of them to B. However, if B is allowed to be arbitrarily large, nothing about the degree of closeness can be concluded. Therefore, we also need the degree to which B is small, i.e., the degree to which every point of B is close to every other point of B w.r.t. R:

$$small_R(B) = \inf_{p\in\mathbb{R}^2} I_W(B(p), \inf_{q\in\mathbb{R}^2} I_W(B(q), R(p,q)))$$

where $R = R_{(\alpha,\beta)}$ for some α in $\mathbb{R}$ and $\beta \geq 0$.

3. GEOGRAPHIC IR

As in the introduction, we assume to have a list of hotels and landmarks with geographical coordinates. In addition however, we construct a list of regions, such as city districts, together with fuzzy footprints representing their extent [5]. To represent spatial descriptions in the framework introduced in the previous section, we first convert natural language sentences to triplets (x, y, r) where x and y are the name of regions, hotels, landmarks, etc., and r is a spatial relation expressed in natural language (e.g., "at the doorstep of"). Initial experimental results with a statistical Named Entity tagger[5] suggest that this can be accomplished very accurately, due to the rather uniform language use. Next, we try to find the class to which x and y belong (e.g., hotel, museum, district, ...), using heuristics — the name of a hotel often begins or ends with "Hotel" — or a gazetteer. Finally, we convert the triplet (x, y, r) to the following 3 assertions (assuming r expresses closeness):

$$small_{R_{(\alpha_1,\beta_1)}}(x) \geq \lambda_1$$
$$small_{R_{(\alpha_2,\beta_2)}}(y) \geq \lambda_2$$
$$C_{R_{(\alpha_3,\beta_3)}}(x,y) \geq \lambda_3$$

To find suitable values for $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$, a data–driven approach can be used. Assume, for example, that we have a large list of pairs (x_i, y_i) $(1 \leq i \leq n)$ for which we find on the web that "x_i lies at the doorstep of y_i", and for which we have the exact distance between x_i and y_i at our disposal. Given a fixed confidence level c in $]0,1]$ (e.g., $c = 0.7$), the values of α_3 and β_3 might be chosen as the smallest values for which

$$\frac{|\{i | 1 \leq i \leq n \text{ and } d(x_i,y_i) \leq \alpha_3\}|}{n} \geq c$$
$$\frac{|\{i | 1 \leq i \leq n \text{ and } d(x_i,y_i) \leq \alpha_3 + \frac{\beta_3}{2}\}|}{n} = 1$$

To find suitable values for $\lambda_1, \lambda_2, \lambda_3$, we first assume that $\lambda_1 = \lambda_2 = \lambda_3 = 1$, i.e., that x and y fully satisfy the relation expressed by r, and that their size corresponds to that of a typical instance of the class to which they belong. The interpretations of λ_1, λ_2 and λ_3 are revised when it turns out that they lead to inconsistency; we omit the details.

The following proposition effectively allows us to deduce knowledge that is only implicitly contained in an initial set of assertions of the form $C_R(x,y) \geq \lambda$, $P(x,y) \geq \lambda$, $NTP_R(x,y) \geq \lambda$ or $small_R(x) \geq \lambda$, where $\lambda \in]0,1]$, $R = R_{(\alpha,\beta)}$ for some $\alpha \in \mathbb{R}$ and $\beta \geq 0$.

PROPOSITION 1. *Let M be the fuzzy spatial relation in the entry of Table 1 on the row corresponding to the fuzzy spatial relation K and the column corresponding to the fuzzy spatial relation L. Furthermore, let $S_a = R_{(\alpha_a,\beta_a)}$, $S_b = R_{(\alpha_b,\beta_b)}$,*

[4]Negative values for α are allowed for technical reasons.

[5]We used Lingpipe, available at http://www.alias-i.com/lingpipe/

Table 1: Transitivity table for fuzzy spatial relations

K \ L	C_{R_2}	P	P^{-1}	NTP_{R_2}	$NTP^{-1}_{R_2}$
C_{R_1}	$C_{R_1 \circ S_b \circ R_2}$	C_{R_1}	$C_{R_1 \circ S_b}$	$NTP_{S_a \lhd (R_1 \lhd R_2)}$	$C_{(R_2 \lhd S_b) \circ R_1}$
P	$C_{S_b \circ R_2}$	P	C_{S_b}	NTP_{R_2}	$C_{R_2 \lhd S_b}$
P^{-1}	C_{R_2}	$C_{R_{(0,0)}}$	P^{-1}	$NTP_{S_a \lhd R_2}$	$NTP^{-1}_{R_2}$
NTP_{R_1}	$C_{(R_1 \lhd S_b) \circ R_2}$	NTP_{R_1}	$C_{R_1 \lhd S_b}$	$NTP_{R_1 \circ R_2}$	$C_{R_2 \lhd (R_1 \lhd S_b)}$
$NTP^{-1}_{R_1}$	$NTP^{-1}_{S_c \lhd (R_2 \lhd R_1)}$	$NTP^{-1}_{S_c \lhd R_1}$	$NTP^{-1}_{R_1}$	$NTP_{S_a \lhd R_2}$	$NTP^{-1}_{R_1 \circ R_2}$

$S_c = R_{(\alpha_c,\beta_c)}$, $R_1 = R_{(\alpha_1,\beta_1)}$ and $R_2 = R_{(\alpha_2,\beta_2)}$ for some $\alpha_a, \alpha_b, \alpha_c, \alpha_1, \alpha_2$ in $\mathbb{R}$, and some $\beta_a, \beta_b, \beta_c, \beta_1, \beta_2 \geq 0$. For fuzzy footprints A, B and C, it holds that[6]

$$M(A,C) \geq T_W(small_{S_a}(A), small_{S_b}(B), small_{S_c}(C), K(A,B), L(B,C))$$

For example, using Proposition 1, from

$$C_{R_{(\alpha_1,\beta_1)}}(x,y) \geq \lambda_1$$
$$small_{R_{(\alpha_s,\beta_s)}}(y) \geq \lambda_s$$
$$C_{R_{(\alpha_2,\beta_2)}}(y,z) \geq \lambda_2$$

we can deduce

$$C_{R_{(\alpha_1,\beta_1)} \circ R_{(\alpha_s,\beta_s)} \circ R_{(\alpha_2,\beta_2)}}(x,z) \geq T_W(\lambda_1, \lambda_s, \lambda_2)$$

or using (2), and assuming $\alpha_1 \geq 0$, $\alpha_s \geq 0$ and $\alpha_2 \geq 0$

$$C_{R_{(\alpha_1+\alpha_s+\alpha_2, \max(\beta_1,\beta_s,\beta_2))}}(x,z) \geq T_W(\lambda_1, \lambda_s, \lambda_2)$$

This formula expresses information on the proximity of x and z, interpreted in terms of the derived resemblance relation.

4. CONCLUSIONS

We have introduced a framework to represent vague spatial information, and provided a transitivity table to support reasoning. Furthermore, we have sketched how this formalism can be used to increase the coverage of a focused IR system, relying on natural language descriptions of spatial constraints when address extraction or geocoding fails.

5. REFERENCES

[1] W. Bandler and L. J. Kohout. Fuzzy relational products as a tool for analysis and synthesis of the behaviour of complex natural and artificial systems. In S. K. Wang and P. P. Chang, editors, *Fuzzy Sets: Theory and Application to Policy Analysis and Information Systems*, pages 341–367. Plenum Press, New York and London, 1980.

[2] M. De Cock and E. E. Kerre. On (un)suitable fuzzy relations to model approximate equality. *Fuzzy Sets and Systems*, 133:137–153, 2003.

[3] D. A. Randell, Z. Cui, and A. G. Cohn. A spatial logic based on regions and connection. In *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, pages 165–176, 1992.

[4] S. Schockaert, C. Cornelis, M. De Cock, and E. E. Kerre. Fuzzy spatial relations between vague regions. In *Proceedings of the 3rd IEEE Conference on Intelligent Systems*, to appear.

[5] S. Schockaert, M. De Cock, and E. E. Kerre. Automatic acquisition of fuzzy footprints. In *Proceedings of SeBGIS 2005, LNCS 3762*, pages 1077–1086, 2005.

[6] We write $T_W(a,b,c)$ as an abbreviation for $T_W(a, T_W(b,c))$ or $T_W(T_W(a,b),c)$. Note that this is not ambiguous, since for all a, b and c in $[0,1]$, it holds that $T_W(a, T_W(b,c)) = T_W(T_W(a,b),c)$.