User Interface Design for Semantic Query Expansion in Geo-data Repositories

Hartwig H. HOCHMAIR and Jennifer FU

Abstract

Semantic query expansion is the process of supplementing a user query with additional terms that interpret and extend the user's information needs. This work presents the results of an empirical study that investigates user preferences for different designs of user interfaces that provide semantic query expansion for data search from geo-data repositories. The study assesses further whether it is possible to map qualitative gradations of semantic relatedness between geographic key terms to ranges of numerical similarity values.

1 Introduction

Geo-data repositories contain information (e.g., documents, datasets, maps, images, biographical information) that are spatially related to geographic locations through placenames (i.e., toponyms) and place codes (e.g., postal codes) or through geospatial referencing (e.g., longitude and latitude coordinates). Metadata provide information about a data file, and typically describe thematic, geographic, and data process characteristics. Metadata are the foundation for a search engine. Multi-modal search interfaces typically provide search options, such as thematic or geographic keywords, interactive mapping activated spatial locator, temporal, and format, etc. This work intends to focus on the semantic query expansion of thematic keyword terms in a geo-data repository, and is motivated by the objectives specified in the Amazon Andean GIS Web Portal (AAGWP) project[1], which are to build a user-friendly GIS clearinghouse to acquire, harvest and publish heterogeneous data and documents, and to facilitate the retrieval of geo-spatial and environmental information.

A recent study has revealed that GIS users, particularly students, experience significant problems in finding and retrieving GIS data both from geo-data warehouses and geo-data clearinghouses (HOCHMAIR 2005), which points out the disconnection between the user interface design and user perception. One portion of problems identified is on the semantic level, namely that a user-entered search term is more specific or more general than the term listed in the metadata set (e.g., water body vs. lake). A solution is to provide a list of controlled vocabularies that the user can pick from, and/or to provide an additional searchable thesaurus browsing tool, such as the NBII thesaurus[2]. Query expansion

[1] http://aagwp.fiu.edu/

[2] Thesaurus of the National Biological Information Infrastructure: http://159.189.176.6/SearchNBIIThesaurus/

automatically supplements entered search terms to increase the chance for the user to find useful data sets. BINDING & TUDHOPE (2004) present a user interface that allows the user to search a thesaurus of artifacts in a museum, and to vary the degree of query expansion through a coarse-grained radio button control. Automated query expansion using a single thesaurus or a combination of thesauri has been target of research for decades in information retrieval (MANDALA et al. 1999). DE COCK & CORNELIS (2005) introduce a query expansion method for free text queries consisting of more than one term. Their approach uses fuzzy rough set theory and takes into account the relevance of new added terms for the query as a whole, as opposed to methods that query key terms individually. Despite the rich repertoire of query expansion algorithms for the WWW and e-commerce applications, only few user interfaces of geo-data repositories provide semantic query expansion functions. One example is the Geosciences Network project interface[3]. It allows the user to browse for a concept (e.g., *plant*) within various provided ontologies (e.g., *biosphere*), and to select a relation between the searched spatial data sets and the concept (e.g., *is related to*). Similarly, little literature can be found on user interface design for automated query expansion. One example is an article by KOENEMANN & BELKIN (1996) that describes user interface design with respect to four different feedback mechanisms, all of which are used to augment the original query. A larger body of literature exists however on user interface design for query formulation (YOUNG & SHNEIDERMAN 1993), and visualization and exploration of search results (SHNEIDERMAN & PLAISANT 2004).

In this paper we investigate through an empirical study with paper questionnaires, which type of interactive elements a user prefers for triggering and specifying semantic query expansion in geo-data repositories (question 1 and 2). Further we examine whether qualitative gradations of semantic similarity, such as “little related” can be assigned to a range of normalized similarity values (question 3) and thus be used in automated query expansion. 14 out of the 20 voluntary participants were graduate students of the geography graduate program at St. Cloud State University, six participants were employees at the GIS-RS Center at Florida International University. Ages ranged from 23 to 50 years (median = 28).

2 Question 1: Preferred Gradations

2.1 Questionnaire Setup and Task

Question 1 addresses the user's comfort with different numbers of gradations for semantic query expansion offered in the user interface. This questionnaire assumes that the query expansion algorithm provides a continuous scaling of relatedness from 0 to 100% between the entered search term and added search terms. A request for higher similarity would return a smaller number of retrieved data sets from the data repository. The finest gradation allows the user to set any similarity value between 0 and 100%, whereas a coarse-grained gradation reduces the user’s number of choices, i.e., the user’s degree of freedom, and pre-classifies the choices into similarity ranges. The 9 designs presented in the questionnaire (Fig. 1) reach from coarse-grained to fine-grained interactive control elements. The designs

[3] http://geongrid.org/

include a varying number of radio buttons (design a-g) and slider bars (h, i). The slider bars provide the highest degree of freedom. We hypothesize that participants prefer the most fine-grained design, as it allows the user to set precisely the desired grade of semantic relatedness between the entered search term and supplemented search terms.

In an introductory text participants were asked to imagine that they had already typed in a thematic search term, such as "geology", and that the application would be able to find data sets related to the specified term. The task of the participants was to rank their top three favorite designs from the list with numeric numbers (1...best design, 2...second best design, 3...third best). No information about the data collection was provided to the users.

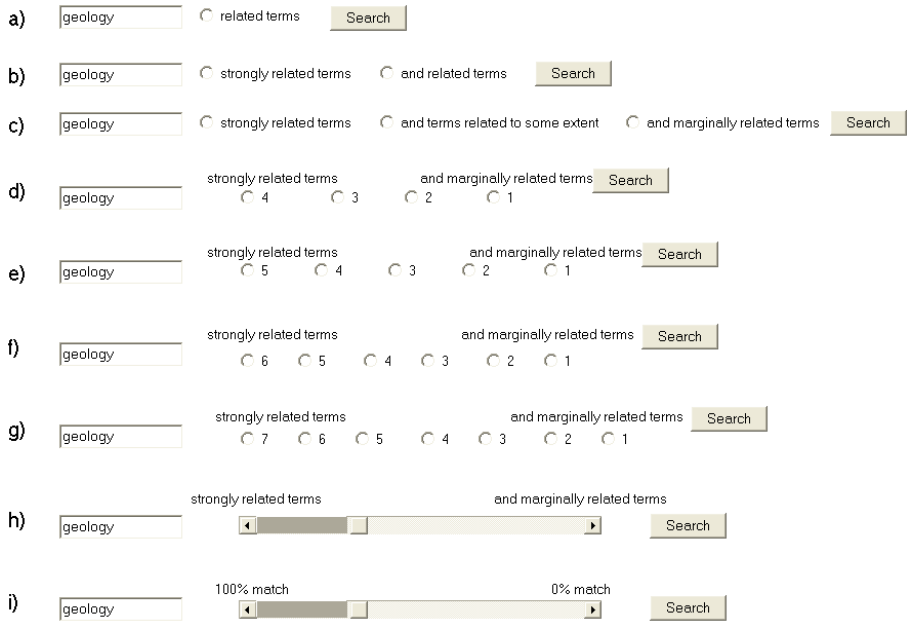


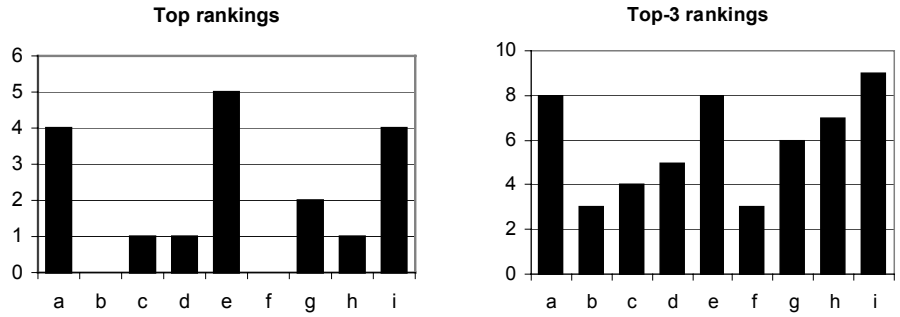
Fig. 1: User interfaces with varying granularity of semantic query expansion

2.2 Results

18 out of 20 participants completed this task correctly, which gave 18 returns for best, second-best, and third-best design out of nine design suggestions. The results indicate that there is no single favorite, but that there are three similarly preferred designs (see Fig. 2a), namely designs *a*, *e*, and *i*. Fig. 2b shows how often each design was ranked among the top-3, which confirms the preference for these designs. Only a small set of the observed preferential differences was significant.

Design *a* is the simplest of all designs, as the user does not need to make any decision on the grade of requested similarity whatsoever. This gives a possible explanation for the high preference for that design. The high rating of design *e* was unexpected, as it seems to be

similar to several other relatively coarse-grained designs in the questionnaire. A possible explanation is that users are familiar with a five-tiered grading scheme from other applications or areas (e.g., the 5-tiered grading scheme in educational systems). Design *i* allows the user to set the semantic relatedness on a continuous scale using a slider bar, which is a possible explanation for the high acceptance rate of this design. The latter characteristics are also true for design *h*. The only difference between *h* and *i* is that *h* uses text labels instead of percentage numbers for describing the degree of match.



(a) (b)

Fig. 2: User preferences for designs with different granularity for query expansion

To check the statistical significance of differences in the user preferences, we converted preference between any two designs to a binary measure because the data are ordinal and not ratio. This was done for all possible 36 design combinations (a-b, a-c, a-d, …, b-c, b-d, …, i-h) for all participants. Two columns were created for each pair of compared designs. If design x was preferred over y by a participant, the x column was assigned value 1 and the y column value 0—or the other way round. If none of the compared designs was listed within the top-3 rankings for that participant, this pair was excluded from the analysis for that participant due to missing values of preference. Then a sign test for two related samples was performed for all design combinations. Tab. 1 shows those design combinations where preferential differences were found to be significant or showed a statistical trend. The $\succ$ symbols means "is preferred to". Only two out of the three favorite designs, namely *a* and *i*, fall into that class. Although design *e* was most often ranked as best design in the patterns of Fig. 1 and Fig. 2, its preference over any other designs is not significant at the 90% significance level.

Tab. 1: Significance level of difference in preference for selected design combinations

	a $\succ$ b	**a $\succ$ c**	**i $\succ$ b**	**i $\succ$ f**
σ (2-tailed)	*.039**	*.039**	*.065*$^{+}$	*.065*$^{+}$

* significant difference of preference between groups ($p < 0.05$)
$^{+}$ statistical trend for difference of preference between groups ($p < 0.10$)

In question 1 we assume that the function which computes similarity values provides a continuous scale between 0 and 100%. However, some algorithms might provide only classed measures of similarity, such as a binary “is similar” and “is not similar”. In such case, a five-tiered or continuous preference scale would mislead the user by pretending a non-existing high fineness of granularity of the underlying similarity function. To avoid this, the provided level of granularity in the user interface should at most be as fine-grained as the underlying system function. Alternatively, the application could indicate on the interactive element, e.g., on the slider bar, how fine-grained the algorithm of the system works, or what the values of semantic similarity for existing data sets look like. This way the user could assess which minimum amount of change on the slider bar would affect the query results. An automated prefetching of potentially relevant datasets, once the keyword has been typed in or selected, would be one method to provide such information about the semantic distance of available data sets.

3 Question 2: Information design

3.1 Questionnaire Setup and Task

Question 2 addresses the user's need for context information before setting the requested value for semantic relatedness in query expansion. We hypothesize that users of geo-data search tools feel more comfortable if they are provided with some concept samples that demonstrate which kind of key terms would be automatically supplemented to the query by the system. The questionnaire contains six designs which can be grouped into three pairs (see Fig. 3). Each pair contains two designs with a shared basic functionality, where the first pair uses radio buttons, the second pair uses slider bars with qualitative labels of relatedness, and the third pair uses slider bars with numerical labels of relatedness. The first design in each pair (*a*, *c*, *e*) does not, whereas the second design in each pair (*b*, *d*, *f*) does provide sample key terms for three levels of similarity. Participants were asked to rank each of the six designs with numeric numbers (1...most preferred, 6...least preferred).

3.2 Results

15 out of 20 participants completed this task correctly. The results suggest some patterns of preferences. However, differences of preferences were not found to be significant, neither between the three basic designs (i.e., the pairs), nor within groups (i.e., design without vs. designs with sample key term). Designs with samples were ranked best 8 times, whereas designs without samples were ranked best 7 times. In the first two groups (a-b and c-d) the median of ranks is smaller for designs that use a sample concept (Fig. 4a). Fig. 4b visualizes the number of times each of the six designs was ranked best or second best among the 15 participants. The results suggests that designs with a sample key term appear more often in the top-2 rankings than designs without a sample key term.

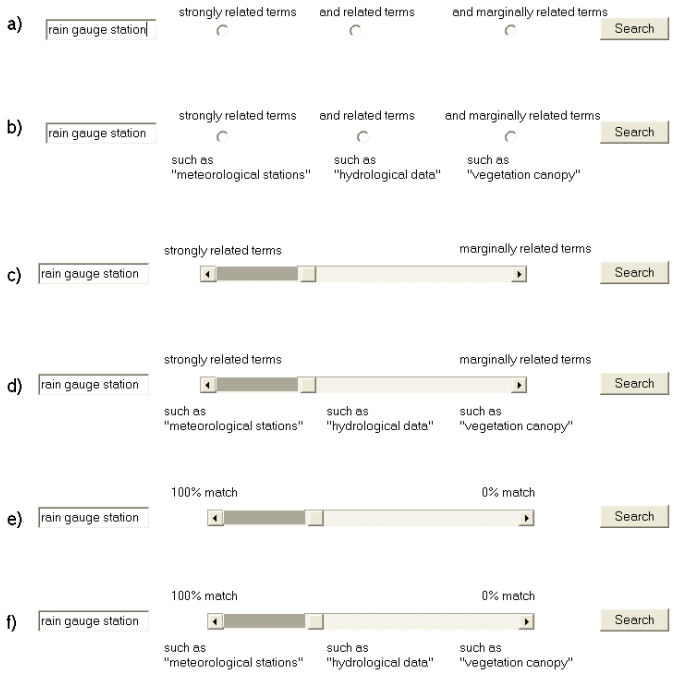
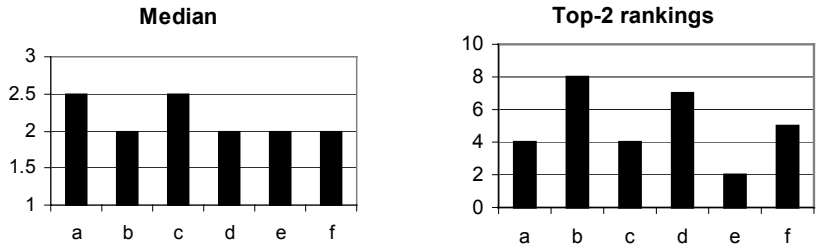


Fig. 3: User interfaces with varying granularity of semantic query expansion



(a) (b)

Fig. 4: User preferences for three different designs with and without concept samples

4 Question 3: Classification of Semantic Relatedness

4.1 Similarity Matrix

In order to provide automated query expansion on keywords, the application needs a distances measure between geographic concepts. Automated methods for assessing the semantic distance between concepts include counting shared and distinguishing features

between key terms based on a thesaurus or an ontology (RODRIGUEZ & EGENHOFER 2004), or comparing the overlap of words used in definitions for concepts in online dictionaries (GREFENSTETTE 1993). For a small number of concepts, similarity measures can be directly assessed through questionnaires. Semantic similarity values can be stored in a similarity matrix that relates all concepts of the database to each other. Normalized similarity values range from S = 0 to S = 1. If the system function provides a continuous scale for similarity values, numerous classification schemes for semantic relatedness can be implemented in the user interface, each realizing a different granularity (compare Fig. 1). It might be helpful for the user to present numerical similarity values in a more tangible way by expressing a range of similarity values with a qualitative term. For example, telling the application to search for "related" and "somewhat related" terms might be more intuitive than specifying a numerical range of 1.0 to 0.4. Question 3 examines whether qualitative gradations of relatedness can be mapped to intervals of numerical similarity values.

For this task, we re-used a similarity matrix for 24 hydrology-related key terms found in a previous study (HOCHMAIR 2006). Cells in the reference matrix (Fig. 5) denote averaged similarity values between pairs of key terms estimated by the 28 participants of that study.

	meterological stations	rain gauge station	landscape	foothills	rivers	hydrological data	meterological data	land cover	soils	topography	geology	watersheds	water sampling station	roads
meterological stations	1.00	0.78	0.20	0.06	0.17	0.44	0.98	0.13	0.12	0.38	0.13	0.30	0.42	…
rain gauge station	0.87	1.00	0.22	0.10	0.30	0.62	0.80	0.22	0.15	0.45	0.13	0.58	0.60	
landscape	0.22	0.08	1.00	0.32	0.53	0.52	0.22	0.60	0.60	0.70	0.52	0.65	0.47	
foothills	0.05	0.12	0.48	1.00	0.10	0.22	0.33	0.23	0.28	0.85	0.57	0.18	0.15	
rivers	0.02	0.14	0.84	0.22	1.00	0.84	0.24	0.32	0.35	0.58	0.48	0.90	0.72	
hydrological data	0.55	0.32	0.64	0.20	0.97	1.00	0.63	0.18	0.22	0.50	0.23	0.80	0.70	
meterological data	0.97	0.76	0.32	0.32	0.35	0.56	1.00	0.55	0.13	0.38	0.12	0.58	0.45	
..	…							…						

Fig. 5: Part of the used similarity matrix for hydrology-related key terms

4.2 Questionnaire Setup and Task

To test for a potential correlation between qualitative terms of relatedness and numerical ranges of similarity values, we created three tasks for participants. For each task we selected the same set of 15 keywords out of the 24 available key terms in the matrix (Fig. 5). The set was chosen in a way that the similarity values between an arbitrary chosen reference keyword, *watershed* in this case, and the other 14 keywords were equally distributed between 0 to 1 according to the reference matrix (Fig. 5). Selected terms included, for example, *catchment* (S = 0.88) or *foothills* (S = 0.18), meaning that *catchment* is more related to *watershed*, than *foothills* is to *watershed*. For task 1 participants were asked to state for each of the 15 terms whether it was related (R) or not related (N) to *watershed*, for task 2 whether it was strongly related (S), related (R), or not related (N) to *watershed*, and for task 3 whether it was strongly related (S), related to some extent (E), little related (L), or not related (N) to *watershed*. Participants made their statements by

checking a corresponding box in the questionnaire. Fig. 6 shows part of the questionnaire for the three tasks. On the handed-out version, the three tasks were answered on separate pages.

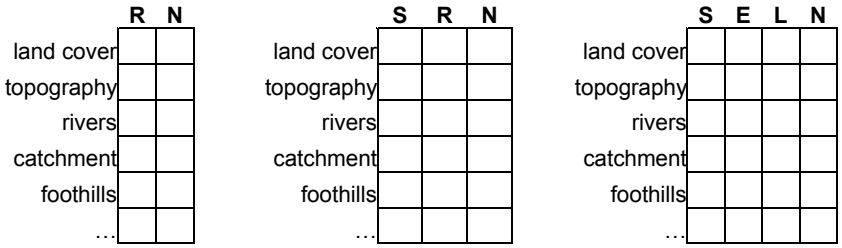


Fig. 6: Questionnaire 3: Examining qualitative gradations of relatedness between *watershed* and other 15 keywords

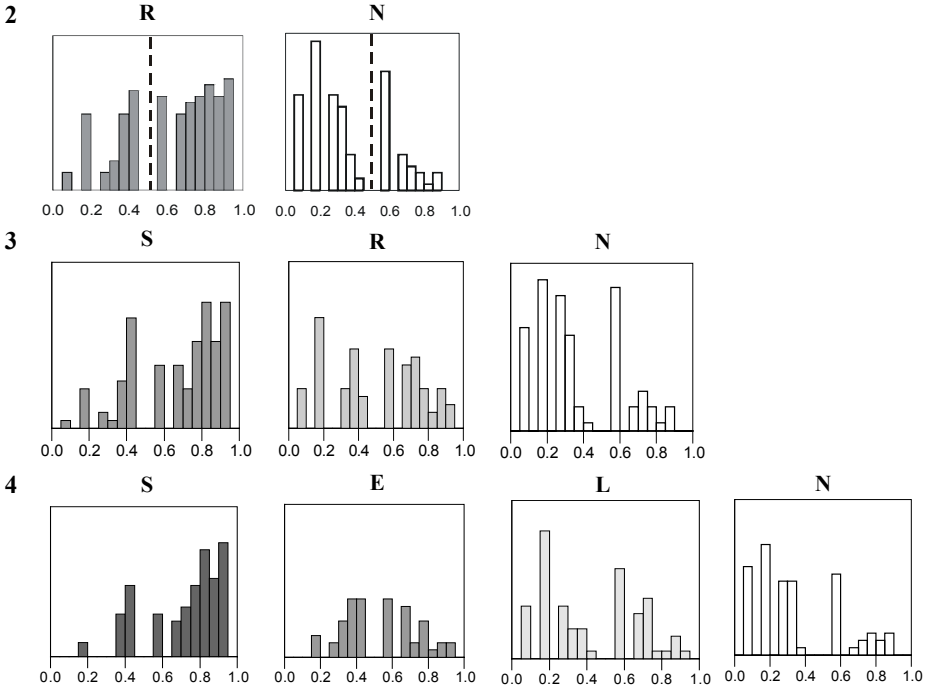
4.3 Results

19 out of 20 participants completed this task correctly. For each task we counted how often each of the 15 terms was assigned to either of the provided classes (R, N, S, E, L). As each concept has a similarity value with *watershed* in the reference matrix, we can create a histogram that visualizes how often a key term of similarity *S* was assigned to a class.

The three rows in Tab. 2 visualize the histograms found for the three classification tasks. Although no clear cut points for assigned similarity values in the individual classes can be identified, we observe the tendency that concepts assigned to classes of higher relatedness share generally higher numerical similarity values with *watershed* than this is the case for concepts assigned to the "little" or "not related" classes. Thus, histograms for classes of higher relatedness are skewed left. If a user specifies the requested semantic relatedness in her query with a qualitative term, the automated query expansion algorithm will—after searching for concepts within the corresponding range of numeric similarity values—yield satisfactory results.

Boundaries of numerical class ranges vary among participants. Sometimes the same concept is assigned to different classes by different participants. For example, both the "related" and "not related" class contain concepts with a similarity value of 0.2 (Tab. 2, first row). Setting crisp class ranges will cause errors of omission and commission in query expansion. Take for example the 2-tiered classification and let us set the class boundary between related and not related terms at S = 0.5 (indicated in Tab. 2 with dashed lines). If the user now requests a search for data sets related to *watershed*, query expansion would omit search for all terms located left of the dashed line in the R-class. This is an error of omission with respect to those users who consider the concepts left of the dashed line in the R-class as related to *watershed*. The query expansion algorithm would search for concepts with an S > 0.5, and therefore also include search for terms that are located right of the dashed line in the N-class. This is an error of commission with respect to those users who consider the concepts right from the dashed line in the N-class as unrelated to *watershed*.

Tab. 2: Histograms for concepts assigned to different classes of qualitative relatedness



Tab. 3: Median of similarity values for concepts assigned to classes of graded relatedness

Classification	**2-tiered**		**3-tiered**			**4-tiered**			
	R	**N**	**S**	**R**	**N**	**S**	**E**	**L**	**N**
Median	.71	.35	.77	.58	.33	.77	.58	.38	.30
σ (2-tailed)	*.000**		*.000**	*.000**		*.000**	*.008**	*.121*	

* difference between groups is significant at the 0.01 level (2-tailed)

Statistical analysis (Mann-Whitney U test) shows that differences between the medians of similarity values assigned to the qualitative classes are statistically significant (Tab. 3) except for the "little related" and "not related" classes in the 4-tiered classification scheme.

5 Summary and Outlook

The first two studies identified preferred user interface features with respect to query expansion. The assumption was that the system would use a function that creates a continuous range of similarity values between the entered search term and search terms

related to other data sets in the data base. Users were not provided with information about the data collection. The results of questions 1 and 2 identified some preferences in user interface design. Generally, participants tend to prefer simple designs with a mere activation function for query expansion, a five tiered classification of semantic relatedness, and a function for setting a similarity value on a continuous scale. Users prefer to be provided with some sample concepts that demonstrate the meaning of qualitative or numerical similarity measures. The third study showed that qualitative gradations of relatedness can be mapped to ranges of numerical similarity values, yet causing errors of omission and commission. For future work we will extend the assessment of user preferences to a system which is based on an iterative cycle considering user feedback, and that provides information about the structure of the data sets in the data collection.

Bibliography

Binding, C. & D. Tudhope (2004). *KOS at your Service: Programmatic Access to Knowledge Organisation Systems*. Journal of Digital Information, *4* (4).

De Cock, M. & C. Cornelis (2005). *Fuzzy Rough Set Based Web Query Expansion*. In Proceedings of Rough Sets and Soft Computing in Intelligent Agent and Web Technology, International Workshop at WIIAT2005 (pp. 9-16).

Grefenstette, G. (1993). *Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches*. ACL'93 workshop on the Acquisition of Lexical Knowledge from Text.

Hochmair, H. H. (2005). *Ontology Matching for Spatial Data Retrieval from Internet Portals*. In M. A. Rodríguez, I. F. Cruz, S. Levashkin & M. J. Egenhofer (Eds.), First International Conference on Geospatial Semantics (GeoS 2005) (LNCS 3799, pp. 166-182). Berlin: Springer.

Hochmair, H. H. (2006). *Filling the Gaps in Keyword-based Query Expansion for Geodata Retrieval*. In W. Kainz, A. Riedl & G. Elmes (Eds.), Spatial Data Handling. Berlin: Springer.

Koenemann, J. & N. J. Belkin (1996). *A case for interaction: A Study of interactive information retrieval behavior and effectiveness*. In M. J. Tauber (Ed.), Human Factors in Computing Systems: CHI '96 Conference Proceedings (pp. 205-212). New York: ACM Press.

Mandala, R., T. Tokunaga, & H. Tanaka (1999). *Combining General Hand-Made and Automatically Constructed Thesauri for Query Expansion in Information Retrieval*. In D. Thomas (Ed.), Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99) (pp. 920–925). Morgan Kaufmann Publishers.

Rodriguez, M. A. & M. J. Egenhofer (2004). *Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure*. International Journal of Geographic Information Science, *18* (3), 229-256.

Shneiderman, B. & C. Plaisant (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Boston: Pearson Education, Inc.

Young, D. & B. Shneiderman (1993). *A graphical filter flow model for Boolean queries: An implementation and experiment*. Journal of the American Society for Information Science, *44* (6), 327-339.