

SPATIAL KNOWLEDGE EXTRACTION FROM GEOGRAPHICAL DATABASES

An approach based on the Controlled English Query Language Geo-Q and Conceptual Graphs

Marinos KAVOURAS and Sofia KONTAXAKI

ABSTRACT

Conceptualizing spatial relations through natural language seems to be closer to the model humans adopt for spatial reasoning. Controlled languages, which are subsets of natural ones, could be used in order to overcome the difficulties that emanate from the required experience in programming languages. In this paper, we propose a methodology in which questions are expressed in a new controlled query language called Geo-Q, for the purpose of extracting spatial knowledge from geographical databases. In the first step of the methodology, the question is parsed and a direct correspondence between the components of the question formulated in Geo-Q and the concepts/relations which constitute a Conceptual Query Graph is established. In the second step, the resulting Conceptual Query Graph is adequately transformed to an SQL query and the execution of the SQL query finally leads to the answer.

KEYWORDS

Spatial Knowledge, Query Language, Natural Language, Conceptual Graph, SQL

INTRODUCTION

Man-machine communication that matches human reasoning and language is an interesting but complex issue. Such achievement would have obvious advantages. According to Rich [13]: “Computers will not be able to perform many of the tasks people do every day until they, too, share the ability to use language”. The truth is that complete ‘understanding’ of natural human language by the machine remains impossible even today due to the complexity of all its aspects: syntactic, semantic and pragmatic. The term ‘understanding’ stands for the transformation of natural language into another form of knowledge representation that may result to computer’s response, e.g. the execution of a computer program.

In order to partially overcome this difficulty, several approaches have been proposed like (a) the semi-automatic language processing requiring human intervention or (b) the creation of controlled languages that constitute subsets of the natural ones with vocabulary and syntax constraints. Controlled languages can be directly transformed into any kind of structured logical expression like First-Order-Logic, programming languages, etc. In this direction the Common Logic Controlled English (CLCE) has been created which can be easily read since «Anyone who can read ordinary English can read sentences in CLCE with little or no training» [17] and subsequently can be translated into First-Order-Logic expressions. Another example, the Attempto Controlled English which has been created in the University of Zurich as a subset of English natural language, can be used in requirements specification and translated to executable Prolog programs [14].

In this paper, the difficulties of accessing and retrieving geospatial information are investigated and a new methodology is proposed in order to overcome them. More specifically, according to the proposed methodology, questions are expressed in a new controlled query language called Geo-Q, for the purpose of extracting spatial knowledge from databases. Geo-Q questions are automatically transformed into Conceptual Query Graphs, which are subsequently transformed into adequate spatial database queries. The execution of these queries leads to the answer.

INFORMATION RETRIEVAL AND QUESTION ANSWERING CONCERNING SPATIAL DATA

People often express and understand spatial relations through natural language instead of metric measurements [1]. When these relations are described through natural language (spoken or written), they seem to be closer to the model humans adopt for spatial reasoning and therefore the user can conceptualize, compare and identify them more easily [2]. On contrary, when a user submits questions or requests to systems that host geospatial data in terms of quantitative data, measures or coordinates, spatial relations are not so easily conceptualized.

Furthermore, information retrieval from geographical database systems has commonly as prerequisite the knowledge of very specialized language for communication and queries' construction (e.g. SQL). Usually this requires the user to be an experienced programmer. On the opposite, geographical information retrieval from the World Wide Web does not require any familiarization with specialized language since search is executed through the usage of keywords. However, this procedure is not very effective because, as a result of such a query, a very big percentage of the retrieved web pages are not relevant to the information searched. In addition, this retrieval method does not offer any support for deep structures 'hidden' in geospatial data so that the user often miss critical information [4].

In order to overcome the difficulties described above we created the Controlled Query Language Geo-Q. Based on this language, spatial knowledge extraction from geospatial structured and unstructured data collections can be accomplished in both human friendly and effective way. The lexical form of Geo-Q has been enriched with geo-spatial concepts and relations and its grammar has been defined in a way to support geo-spatial reasoning.

In the proposed methodology, questions which are originally submitted in Geo-Q are, in a first step, transformed into Conceptual Query Graphs and, in a second step, Conceptual Query Graphs are transformed into adequate database access statements, depending on the system hosting the geospatial data collections.

CONCEPTUAL GRAPHS

Conceptual Graphs (CGs) are networks of concept and relation nodes [3]. The concept nodes represent entities, attributes or events while the relation nodes identify the kind of relationship between two concept nodes. The nodes have arcs between them that are always directed. The notation used in order to describe CGs may be graphical or linear. In the graphical notation, concepts are drawn as rectangles and relations as circles. In the linear notation, concepts are used as words in brackets, while relations have parentheses.

Conceptual Graphs constitute a visual, advanced knowledge based representation formalism grounded on philosophical, linguistic and object-oriented principles [15, 16] and are considered as the most general and flexible logic notation that can be transformed to any form of logic [18]. CGs have already been used in Text Mining in order to discover knowledge in large data collections [10, 6] but also for Information Retrieval [8, 9]. In addition, CGs have been used to model, index and retrieve images [11]. In geospatial domain, CGs have also found use in the field of geospatial knowledge representation derived from definitions expressed in natural language [9].

QUESTIONS SUPPORTED BY GEO-Q

In this paper, we examine the case of submitting questions to geographical data bases systems in order to extract spatial knowledge. Figure 1 describes the successive question processing and transformation of the originally submitted Geo-Q questions.

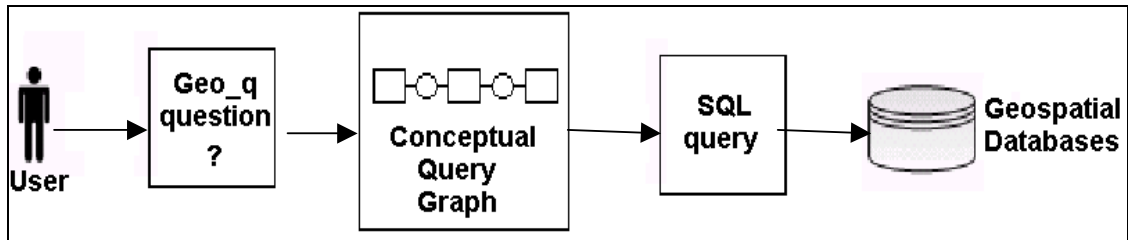


figure 1: Successive question processing and transformation

Variety of questions can be set to geographical database systems. In this version of Geo-Q, we focused on the construction of the most representative ones. These questions types (Q1, Q2 and Q3) are presented in following:

Type Q1: questions that produce answers like YES or NO

- Is polygon A adjacent to polygon B?
- Does polygon A touch polygon B?
- Is point X at the North of polygon A?

Type Q2: questions that return one or more geo-entities as an answer

- Which line intersects polygon A?
- Which polygon is closest to polygon B?
- Which points are at the North of polygon A?

Type Q3: questions that return the attribute's value of a geo-entity

- What is the color of polygon A?
- What are the coordinates of X?
- What is the shape of B?

GEO-Q GRAMMAR

For the construction of the Geo-Q language, the definition of its lexical and syntactical form, that is its grammar, is needed.

Geo-Q lexical form

Language vocabulary consists of:

- a. Words supporting question's construction like: 'which', 'what', 'of', 'and', 'the', 'a', 'an', 'some', 'many', 'any', etc.
- b. Words participating in allowed spatial expressions like: 'At the North', 'At the South', 'Near', 'Between', etc.
- c. Words corresponding to verbs that express spatial relations and which can be used in singular/plural and active/passive voice like: 'intersects', 'intersect', 'intersected by', etc.
- d. Words derived from the spatial data base: tables and fields names.

Moreover, submitted questions are considered as completed only when finishing with symbol '?'.

Geo-Q syntactical form

Geo-Q syntactical form is described by its grammar which has been defined using the Backus-Naur Form [5]. Since current version of Geo-Q is limited to support Q1, Q2 and Q3 types, the main body of questions is defined by the expandable grammar rules:

<question> ::= <question_type> ?

<question_type> ::= <Q1> | <Q2> | <Q3>

In the past, spatial relations have been classified into several types [12]. Current version of Geo-Q supports the use of terms corresponding to a small subset of the topological, ordinal and distance relations that can be used in natural language. Spatial relations are defined by grammatical rules based on expression which use the verb 'be', other verbs and geospatial expressions without any verb in their structure:

<relation_expression> ::= <verb_expr> | <geo_expr>¹

<verb_expr> ::= <active_expr> | <passive_expr>

<active_expr> ::= <plur_verb>² | <sing_verb>³

<passive_expr> ::= <verb_be> | <past_participle>⁴

<verb_be> ::= is | are

Where:

1. geo_expr: allowed geospatial expressions like: 'At the North', 'At the South', 'Near', 'Between', etc.
2. plur_verb: allowed active and plural verb like: 'Touch', 'Intersect', 'Overlap', etc.
3. sing_verb: allowed active and singular verb like: 'Touches', 'Intersects', 'Overlaps', etc.
4. past_participle: allowed expressions using verb in passive voice like: 'Touched by', 'Intersected by', 'Overlapped by', etc.

References to geospatial entities are possible through expressions combining nouns, articles and adjectives. Nouns and adjectives correspond to the terms that define tables and fields in the geographical database. Furthermore, user is allowed to use singular or plural expressions like he would do in natural language. The grammatical rules that describe this kind of expressions are:

<many_geoentities> ::= <geoentity> [[{, <geoentity>}] and <geoentity>]

<geoentity> ::= [<articles>] [<qualifiers>] [entity_type¹] entity_name² [<qualifiers>]

<articles> ::= (the | a | an | some | many | any)

<qualifiers> ::= <qualifier> [[{, <qualifier>}] and <qualifier>]

<qualifier> ::= entity_attribute³

Where:

1. entity_type: the name of a geospatial entity that corresponds to a specific table name in the geographical database. For example: 'Polygon', 'Line', etc.
2. entity_name: the name of the geospatial entity characteristic that corresponds to a specific field name defined as a primary key of a table in the geographical database. For example: 'A', 'B', 'X', etc.

- entity_attribute: the name of the geospatial entity characteristic that corresponds to a specific field name of a table in the geographical database. For example: 'coordinates', 'color', 'shape', etc.

FIRST STEP: GEO-Q QUESTIONS TRANSFORMATION INTO CONCEPTUAL GRAPHS

There is direct correspondence between the components of a question formulated in Geo-Q and the concepts/relations which constitute the Conceptual Query Graph. But, while this approach aims to be a user-friendly one, it requires that the user formulates his questions using terms that could be mapped to database features. To this point, we propose that a database catalogue should be provided by the system to help the user select the correct tables and fields or to define his own aliases. For verbs or other geographical expressions corresponding to some spatial relation between the data, we suppose that a table exists in the database or can be computed and integrated in the database in order to define the relation (e.g. 'Touches').

In the first step of the methodology, the components of the question are mapped to the proper concepts/relations of the Conceptual Query Graph. This process is described for each type of questions in following.

Question type Q1

Q1 type questions examine if a spatial relation among geospatial entities is valid or if specific attributes characterize a set of geospatial entities in order to conclude to a YES or NO answer. The transformation of the Geo-Q question into a Conceptual Query Graph varies depending on the relation expression that is used: verb in active or passive voice, verb 'be' or geospatial expression. Question mark is found in the middle concept of the graph which corresponds either to the relation between geospatial entities or to the verb 'be' between a geospatial entity and a characteristic. In figure 2, alternative scenarios are shown with different line types.

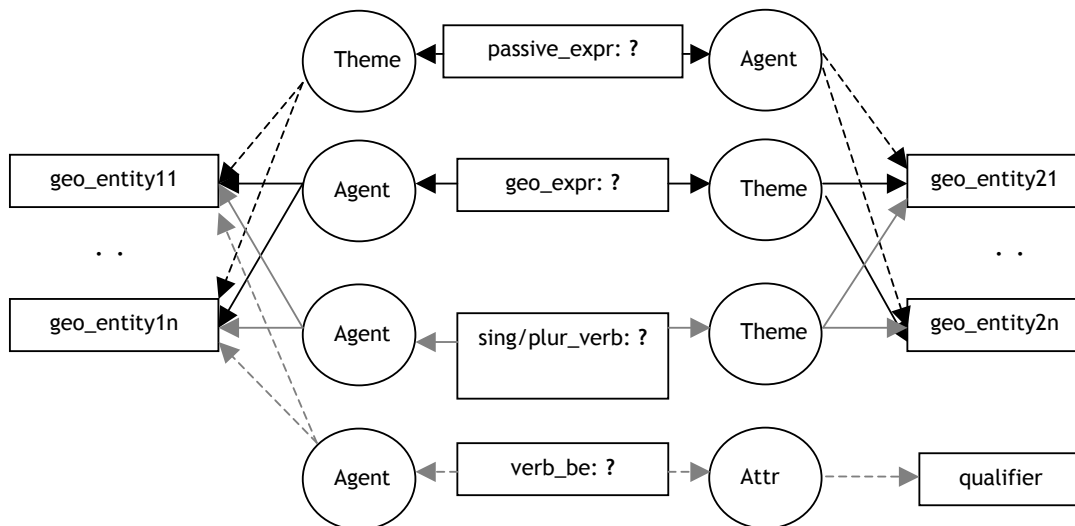


figure 2: Alternative scenarios for the transformation of Q1 type question to a CG query.

The question mark in the Conceptual Graph presented above means that it is a Query Graph.

Question type Q2

Q2 type questions search for the set of geospatial entities that are connected to a second set of geospatial entities through a specific relation. As for the Q1 type questions, the transformation of the Geo-Q question into a Conceptual Query Graph varies according to the expression used to describe the

relation between the two sets of geospatial entities. In this case, the question mark is moved to the concepts found on the left side of the graph which correspond to those the user is looking for. In figure 3, alternative scenarios for Q2 question type are shown with different line types.

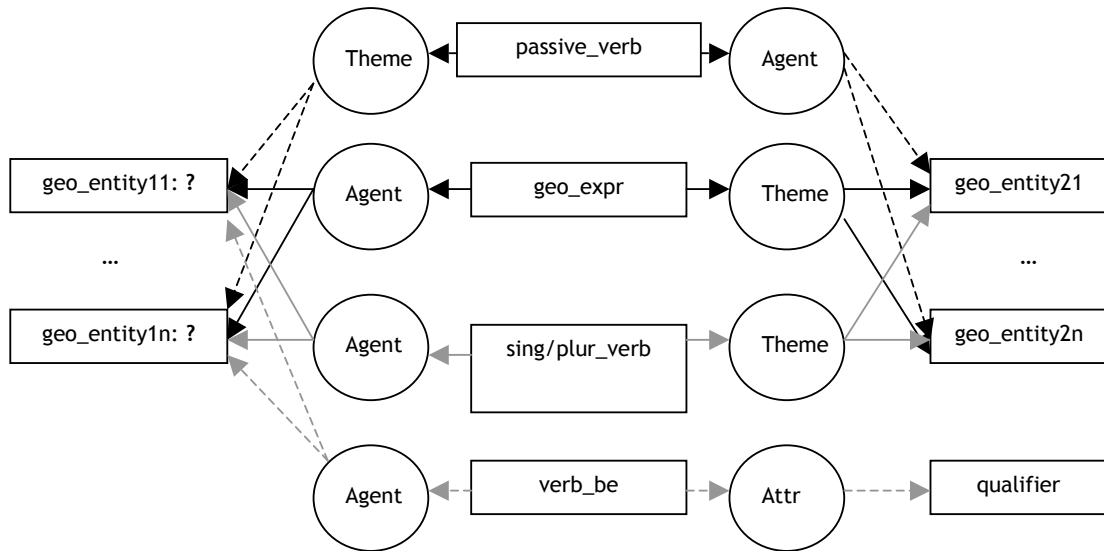


figure 3: Alternative scenarios for the transformation of Q2 type question to a CG query.

Question type Q3

For question type Q3, the characteristic or attribute of one or many geospatial entities is searched. The Conceptual Query Graph of this type of question is figured below (figure 4). Question mark has moved to the concept at the right side of the graph which results to the common characteristic of the entities.

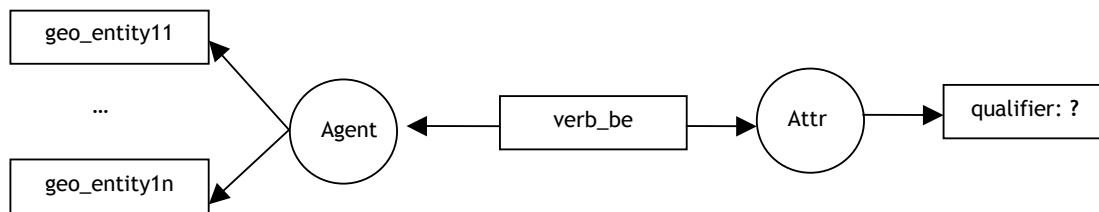


figure 4: Transformation of Q3 type question to a CG query.

SECOND STEP: TRANSFORMATION OF CONCEPTUAL QUERY GRAPH INTO SQL QUERY

The fact that the system 'understands' the question means that it is capable to process and transform it into another form of logic. This is done during the second step of the methodology. After the parsing of Geo-Q questions, the corresponding Conceptual Query Graph is constructed and transformed into adequate SQL query.

For better understanding of this transformation procedure, an example will be examined in following. Suppose we are referring to the geographical database containing the following tables:

Node	X1	X2
1	25	78
2	30	82
3	28	76
4	67	67
5	35	83

table 1: nodes defined in database

Arc	Srce_node	End_node
a	1	2
b	3	2
c	3	1
d	1	4
e	3	4
f	5	5
g	4	2

table 2: arcs defined in database

Poly	Arc_num	Arc_list
A	3	a, d, g
B	3	c, d, e
C	1	f
D	4	b, e, g, f

table 3: polygons defined in database

Graphical representation of the data corresponding to the tables above is shown in figure 5.

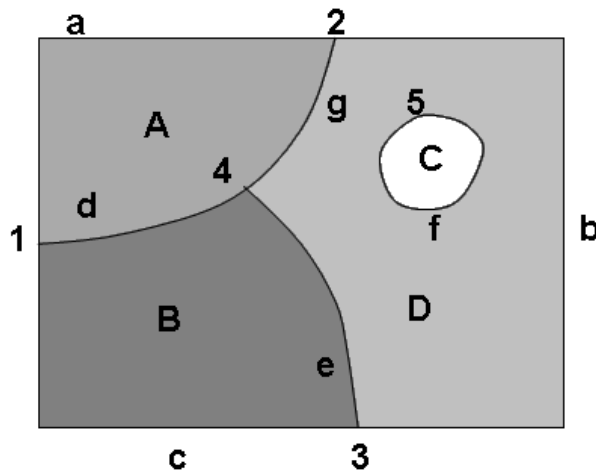


figure 5: Graphical representation of the example data.

Before the second step of the methodology takes place, it is necessary to reorganize the database in terms of Conceptual Graphs formalism. Every table of the geographical database corresponds to a relation in terms of Conceptual Graphs [16]. Consequently, from the structure of the data contained in the geographical database, the following relations can be defined respectively from table1, table2 and table3:

relation Nodes (*x,*y,*z) is (2)

[Node: ?x] -
 (Coord)-> [X1: ?y]
 (Coord)-> [X2: ?z]

relation Arcs (*x,*y,*z) is (3)

[Arc: ?x] -
 (Srce_node)-> [Node: ?y]
 (End_node)-> [Node: ?z]

relation Polys (*x,*y,*z1,...,*zn) is (4)

[Poly: ?x] -
 (Arc_num)-> [Int_num: ?y]
 (Arc_list) -> [Arc: ?z1]
 ...
 (Arc_list) -> [Arc: ?zn]

} Arc_num

In addition, we assume that the concept ‘touches’ is defined as the table describing the polygons that are in contact with others, as shown in table 4:

Poly1	Poly2
A	B
A	D
B	A
B	D
C	D
D	A
D	B
D	C

table 4: the concept ‘touches’ as it is defined as a table in the geospatial database

The table ‘Touches’, represented as a relation in terms of Conceptual Graphs formalism, becomes:

relation Touches (*x,*y) is [Poly1: ?y] <- (Agnt) <- [Touches] -> (Thme) ->[Poly2: ?x] (5)

where:

type Poly1 (*x) is [Poly: ?x] <- (Agnt) <- [Touches] (6)

type Poly2 (*x) is [Poly: ?x] <- (Thme) <- [Touches] (7)

In order to answer a question formulated in the form of a query graph, Sowa [16] states that: “Restructuring a large database is a lengthy process that is sometimes necessary. But to answer a single

question, it is usually faster to restructure the query graph than to restructure the entire database”. In our example, the database is a small one but if it was larger we would just have to restructure the tables participating to the query.

Suppose we are interested in finding the answer to the question: «Which polygon touches A?». The parsing of the question results to a Q2 question type. According to figure 2, ‘polygon’ is mapped to the corresponding concept ‘geo_entity11’, ‘Touches’ to concept ‘sing/plur_verb’ and ‘A’ to concept ‘geo_entity21’. The question is then transformed into the Conceptual Query Graph shown in figure 6.



figure 6: Conceptual Graph query of the question «Which polygon touches A? »:

The symbol {*} indicates that the noun ‘Polygon’ was used in plural form in the initial question. The concepts which constitute the query graph of the question are: ‘polygon’, ‘touches’ and ‘A’. ‘Polygon’ and ‘A’ can be defined as following:

type polygon (*x) is [Poly: ?x] (8)

type A (*x) is [Poly: A] (9)

The Conceptual Query Graph of the figure 6 is then transformed sequentially as it is shown below:

Which polygons touch A?	Example of initial Geo-Q question
⇒ [Polygon: {*}?] <- (Agnt) <- [Touches] -> (Thme) -> [A]	(Conceptual Query Graph, fig.6)
⇒ [Poly: {*}?] <- (Agnt) <- [Touches] -> (Thme) -> [Poly: A]	(implied from (8) and (9))
⇒ [Poly1: {*}?] <- (Touches) -> [Poly2: A] (10)	(implied from (5))

The ‘select’ SQL statement is the combined with the concept that contains the question mark [16], that is [Poly1: {*}?]. The relation (Touches) of the query graph is used as an argument for the ‘from-clause’ and the concept [Poly2:A] is mapped to the ‘where-clause’ which defines the conditions of the select statements [16]. As a result, the Conceptual Query Graph results to the following SQL query (11):

```
Select poly1
  from Touches
  where poly2 = 'A' (11)
```

The execution of the SQL query (11) finally leads to the answer: ‘B, D’.

CONCLUSION AND FURTHER DEVELOPMENTS

Conceptualizing spatial relations through natural language seems to be closer to the model humans adopt for spatial reasoning. Controlled languages, being subsets of the natural one, could be used in order to overcome the difficulties that results from the use of quantitative information and specialized programming languages. In this paper, we show how the controlled language Geo-Q is used to submit questions for the purpose of extracting spatial knowledge from databases based on Conceptual Graphs.

In the future, Qeo-Q grammar will be enriched in order to support a bigger variety of questions, richer vocabulary and more flexible syntax. Our goal is to create a Geospatial Question-Answering System where the capability of submitting questions using the Controlled English language Geo-Q will be encapsulated in order to find accurate and concise answers from heterogeneous geospatial data collections and visualize them in an effective way.

REFERENCES

1. B., Aleman-Meza, C., Halaschek, A., Sheth, I.B., Arpinar, and G., Sannapareddy, SWETO: Large-Scale Semantic Web Test-bed. Intl. Workshop on Ontology in Action, Banff, Canada, June 20-24, 2004.
2. B., Arpinar, A., Sheth, C., Ramakrishnan, L., Usery, M., Azami & M., Kwan. Geospatial Ontology Development and Semantic Analytics. Book Chapter, Handbook of Geographic Information Science, Eds: J. P. Wilson and A. S. Fotheringham, Blackwell Publishing, July 2005 (in print).
3. Conceptual Graph Standard, 2002 NCITS.T2 Committee on Information Interchange and Interpretation. <http://users.bestweb.net/~sowa/cg/cgstand.htm>.
4. Egenhofer, J.M., Toward the Semantic Geospatial Web. In Proceedings of the Tenth ACM International Symposium on Advances in Geographic Information Systems, McLean, Virginia, 2002.
5. Extended Backus-Naur form (EBNF) Syntax. ISO/IEC 14977:1996 standard. Draft document. University of Cambridge, Computer Laboratory. <http://www.cl.cam.ac.uk/~mgk25/iso-14977.pdf>
6. Hensman, S., Construction of Conceptual Graph representation of text. In Proceedings of the Student Research Workshop at HLT-NAACL, Boston, USA, May 2, 2004, pp. 49-54.
7. Karalopoulos, A., Kokla, M., & Kavouras, M., Geographic Knowledge Representation Using Conceptual Graphs. In Proceedings of the 7th AGILE Conference on Geographic Information Science, Science, Crete, Greece, 28 April - 1 May 2004
8. Jiwei Zhong, Haiping Zhu, Jianming Li and Yong Yu, Conceptual Graph Matching for Semantic Search, ICCS 2002: 92-196.
9. Huibers, T., Ounis, I. and Chevallet, J.-P., Conceptual Graph Aboutness. In P.W. Eklund, G. Ellis en G. Mann, editors, Conceptual Structures: Knowledge Representation as Interlingua, 4th International Conference on Conceptual Structures (ICCS'96), volume 1115 of Lecture Notes in Artificial Intelligence, pp 130-144, Sydney, Australia, August, 1996. Springer-Verlag, Berlin.
10. Montes-y-Gómez, M., Gelbukh, A. and López-López, A. Text mining at Detail Level using Conceptual Graphs, 10th International Conference on Conceptual Structures, ICCS 2002, Borovets, Bulgaria, Julio 2002. Lecture Notes in Artificial Intelligence, vol 2393, Springer, 2002.
11. Ounis, I., and Pasca, M., 1998. Modeling, Indexing and Retrieving Images Using Conceptual Graphs. Proceedings of the 9th {DEXA} International Conference on Database and EXpert Systems Applications. Vienna, Austria, 1998. Eds: G. Quirchmayr and E. Schweighofer and T.J.M. Bench-Capon, pp 226-239.
12. Pullar, D. & Egenhofer, M. Toward formal definitions of topological relations among spatial objects. Proceedings of the Third International Symposium on Spatial Data Handling; Sydney, Australia, 1988.
13. Rich, E., 1983. Artificial Intelligence. McGraw-Hill Book Co. ISBN 0-07-052261-8, pp436.
14. Schwitter, R., Fuchs, N.E., & Schwertel, U., Attempto - Controlled English (ACE) for Software

Specifications. Second International Workshop on Controlled Language Applications, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, May 1998.

15. Sowa, J., 1984. Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley. ISBN 05214449004, pp 406.
16. Sowa, J., 2000. Knowledge Representation: Logical, Philosophical and Computational Foundations. Brooks Cole Publishing Co. ISBN 0-534-9496-7, pp594.
17. Sowa, J., Common Logic Controlled English Specifications. <http://www.jfsowa.com/clce/specs.htm>, February, 2004.
18. Sowa, J., Graphics and Languages For the Flexible Modular Framework. <http://www.jfsowa.com/pubs/gal4fmf.htm>, paper presented at the International Conference on Conceptual Structures (ICCS) during the week of July 19 to 23, 2004.

AUTHORS INFORMATION

Marinos KAVOURAS

mkav@survey.ntua.gr
Cartography Laboratory
School of Rural and Surveying Engineering
National Technical University of Athens

Sofia KONTAXAKI

skontax@mail.ntua.gr
Cartography Laboratory
School of Rural and Surveying Engineering
National Technical University of Athens