A Statistical and Structural Approach for Symbol Recognition, Using XML Modelling

Mathieu Delalandre[1], Pierre Héroux[1], Sébastien Adam[1],
Eric Trupin[1], Jean-Marc Ogier[2]

[1] Laboratory PSI, University of Rouen, 76 821 Mont Saint Aignan, France
[2] Laboratory L3I, University of La Rochelle, 17042 La Rochelle, France

Abstract. This paper deals with the problem of symbol recognition in technical document interpretation. We present a system using a statistical and structural approach. This system uses two interpretation levels. In a first level, the system extracts and recognizes the loops of symbols. In the second level, it relies on proximity relations between the loops in order to rebuild loop graphs, and then to recognize the complete symbols. Our aim is to build a generic device, so we have tried to outsource models descriptions and tools parameters. Data manipulated by our system are modelling in XML. This gives the system the ability to interface tools using different communication data structures, and to create graphic representation of process results.

1 Introduction

The current improvements of intranet structures allow large companies to develop internal communications between services. The representation of the heritage of huge companies like a network managers firm is often represented through paper documents, which can be either graphic or textual. As a consequence, the sharing of these kind of information will stay very difficult as long as the storage format will not be digital. This explains the current development of studies concerning the automatic analysis of cartographic or engineering documents, which comes as a result of the growing needs of industries and local groups in the development and use of maps and charts. The aim of the interpretation of technical maps is to make the production of documents easier by proposing a set of steps to transform the paper map into interpreted numerical storage [1][2][3][4]. An important step of this conversion process consists in the recognition of symbols, which often appear on technical documents.

We present in this document a symbol recognition system. It is based on a statistical and structural approach combination. In the second section, we will briefly describe the classical approaches for symbol recognition. Then, we will present our approach. Finally, we will give conclusions and propose some perspectives for our future works.

2   Classical Approaches for Symbol Recognition

Symbols constitute an important informative source on technical documents (geographical maps, utility maps, architectural drawings…). Good states of the art dealing with such a problem can be found in [5][6].

These surveys show that structural approaches are generally chosen for symbol recognition. Such approaches begin with a graphical primitives extraction step. These primitives can be either structural [7][8][9] or statistical-structural [10]. After this first step, primitives and their relations are then represented into a graph, which is used in a process of sub-graph matching in a reference graph. Nowadays, such structural symbol recognition systems are generally efficient for specific applications but cannot be generalized. Only the works of Messmer [11], Schettini [12] and Pasternak [7] can be considered as generic approaches. Indeed, these authors propose generic symbol description tools. Symbols are described by different primitives obtained through the use of low-level operators, and by the association relations between these primitives.

However, some problems are not solved in these systems:

- x Limitations appear when symbols integrate important variability, or when they are represented by elements, which are closed but not connected.
- x Very few works propose a correction step of the primitives extracted by low-level operators. Yet, this point is important, essentially in the case of damaged documents, for which low-level operators can be disrupted by noise.
- x Symbol representations are generally dedicated to existing tool libraries and to specific applications. A system allowing a more global representation of symbols does not exist.

3   Our Statistical and Structural Approach for Symbol Recognition

3.1   Introduction

Our approach may be decomposed into 3 steps:

- x Extraction of loops
- x Extraction of orientation invariant features and statistical recognition of loops
- x Reconstruction of loop graphs, and structural symbol recognition

The system relies on proximity relations between the loops in order to recognize the symbols. Our aim is to build a generic device. So, we have tried to outsource from algorithms model descriptions and tool parameters. In this way, this system is evolutionary and can be used in practice for different applications. Until now, it has been exploited only for symbol recognition on France telecom (a French telecommunication operator) utility maps, and for meteorological symbol recognition.

The France Telecom symbols represent technical equipments permitting connections on telephonic network: Concentration Points "Point de Concentration" and room

"*chambre*". They are composed of a variable number of loops belonging to 5 different classes. The Fig. 1.a shows these symbols. On the top one can see from left to right: « *chambre, PC paires sur bornes, PC paires sur appui* ». Just below one can see the 5 loops classes named from left to right: « *chambre, ellipse, triangle, cercle, portion* ».

The meteorological symbols represent cloud cover rate. They are composed of a variable number of loops belonging to 4 different classes. These symbols are shown on Fig.1.b. On the top one can see from left to right: « *aucun nuage, 2/10 à 3/10 du ciel couvert, 4/10 du ciel couvert, 5/10 du ciel couvert, 6/10 du ciel couvert, 6/10 à 7/10 du ciel couvert, ciel obscurci* ». Just below one can see the 4 loops classes named from left to right: « *cercle_25, cercle_50, cercle_75, cercle_100* ».
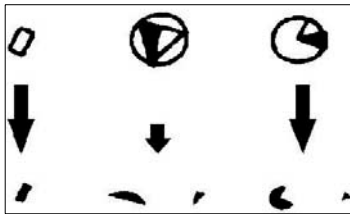
**Fig. 1.a)** France Telecom utility map symbols and their loops
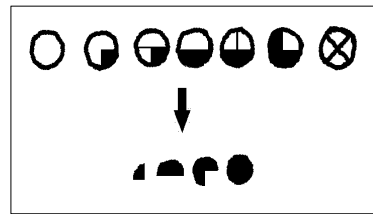
**Fig. 1.b)** Meteorological symbols and their loops

In the following, we will present successively each of the 3 processing steps. We will present succinctly the two first steps (loop extraction and classification) to develop more extensively the structural recognition step. Then, we will present in subsection 3.5 the strategy used for the application and the obtained results. Finally, we will present in subsection 3.6 XML use in our system.

3.2   Extraction of Loops from Symbols

An image of loops is obtained through the application of a classical connected components extraction on image. The Fig. 2.a and 2.b show a part of France Telecom utility map as well as the loop extraction result.
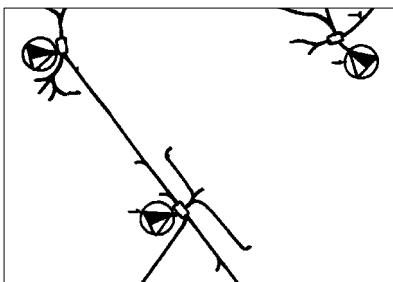
Fig. 2.a) A part of France Telecom utility map
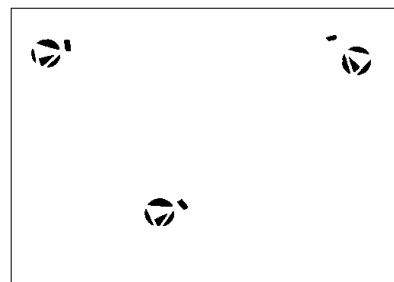
Fig. 2.b) Result after extraction of loops

3.3   Extraction of Orientation Invariant Features and Statistical Recognition of Loops

This processing step constitutes the statistical approach of our system. A feature vector is extracted for each loop on the image, using three outstanding and complementary tools: the Zernike moments, the Fourier-Mellin invariants, and the circular probes [13]. These features enable to constitute a vector set describing loops. This description is invariant to scale and orientation changes. We have constituted a test and a training set using France Télécom utility map loops, with a size of fifty loops each. Then, with the help of the k nearest neighbors (knn, with k=1) classifier, using Euclidian distance, we got recognition results that are presented in the Fig.3 for each feature extraction tool.

| Characteristics extraction tools	Recognition rates
Zernike moments	97.77 %
Fourier-Mellin invariants	86.66 %
Circular probes	86.66 %

Fig. 3. Results of loop recognition

These results show that Zernike moments are the most adapted for the recognition of this loop type. Results are quite the same for meteorological symbols. Of course, these recognition rates computed on a test set of weak size are not representative of a real problem, but they indicate promising perspectives of recognition on test sets of big size.

3.4 Reconstruction of Loop Graphs, then Structural Symbol Recognition

This processing step constitutes the structural approach of our system. It may be divided into two steps.

The first step is a model reconstruction step in the sequential processing between the statistical classifier and the structural classifier showed in the Fig. 4.

| Statistical classifier	→	Models reconstruction	→	Structural classifier
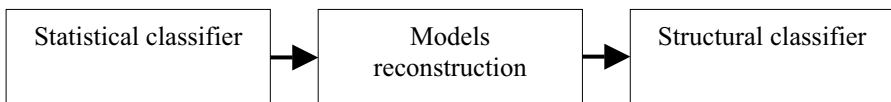
Fig. 4. Sequential processing of statistical and structural classification

The models reconstruction tool we used rebuilds some graphs under connection and/or distance constraints. This reconstruction uses the results of loop statistical recognition. The distance constraint permits to control inter-connection of the graphs corresponding to the symbols. It is thus possible to create a graph in which image loops are completely inter-connected, or to isolate each of the symbols of the image.

These connections constraints are defined according to the nature of loops. The maximal connections number is specified for each loop class. The symbol description is *a priori* considered in model reconstruction strategy. The Fig. 5.a shows an example of France Telecom utility map. Model reconstruction of this map is achieved by using constraints concerning connection and distance constraints. The distance constraint permits to detect all the 4 graphs of loops. The Fig. 5.b shows the graphic representation of the model reconstruction of the symbol located below on the right of the map. This graphic representation uses the information obtained from the statistical classification, from the model reconstruction, from a step of contours detection and polygonisation of the loop image. Here, we have a 4-connections constraint for the loop "*triangle*", and a 1-connection constraint for the loops "*ellipse*" and for the loop "*chambre*".
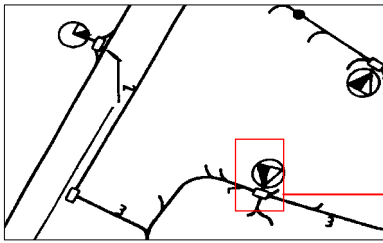
Fig. 5.a) Example of France Telecom utility map
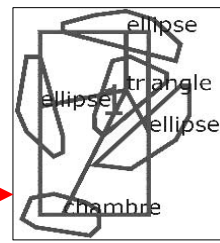
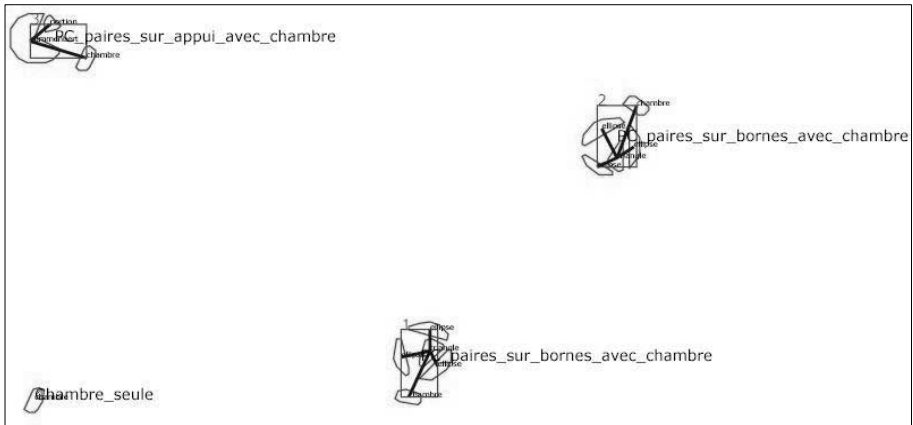**Fig. 5.b)** Graphic representation of the model reconstruction of the right low symbol

The second step is the structural recognition of symbols. It consists in submitting the graphs obtained from the model reconstruction phase to the structural classifier. Our graph matching tool [14] allows a graph edges and nodes typing possibility (integer, float, character, string, object). It permits to compute similarity criterion between graphs, based on the overlap between a candidate graph and a model graph. This overlap corresponds to their common sub-graph. This common sub-graph is searched in three times by matching candidate graph and model graph. In a first time, a filtering step aims at suppressing in the two graphs the nodes and their edges unable to be matched. This concerns the nodes whose label is not common in the two graphs nodes lists. This step has for purpose to reduce the algorithm temporal complexity. In a second time, a research of matching edges is done. The edges are matched if they are equal and if their extremities are equal. The nodes corresponding to the edges extremities are also matched during the edge matching. In a third time, the remaining nodes are matched. Two similarity criteria can finally be computed according to the number of common elements, either on the nodes (1), either on the edges (2). In these equations, n1, n2, and nc and e1, e2, and ec respectively represent the number of nodes and edges of the graph 1, the graph 2, and the common sub-graph.

$$c(g1, g2) = \frac{n1 \times n2}{nc^2} \leq 1 \quad (1) \qquad c(g1, g2) = \frac{e1 \times e2}{ec^2} \leq 1 \quad (2)$$

It is possible to control the combination of the results obtained out of (1) and (2) by average or variance computation. The similarity criterion computation tool is also parameterized in order to take or not into account the types and their attributes. It can be done on the nodes and/or on the edges. For example, it is possible to compute the graph matching taking into account the graph topology, or the graph types. Finally, it is possible to combine the whole similarity criteria obtained by the computation of weighted average, in order to get global similarity criterion.

The Fig. 6 is a graphic representation of a model extraction (under connection and distance constraints) and a structural recognition, the loop graphs have been submitted to the structural classifier with a similarity computation taking into account the types and their attributes based on an average of (1) and of (2).


Fig. 6. Graphic representation of treatment steps of the Fig. 5.a

Fig. 7.a shows a meteorological symbol image. The Fig 7.b shows a graphic representation (of model extraction "only under distance constraint" and a structural recognition) superposition with meteorological symbol image. The loop graphs have been submitted to the structural classifier with a similarity computation taking into account the types and their attributes only based on (1).

Fig. 7. a) Example of meteorological symbol image
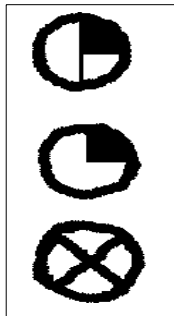
Fig. 7.b) Graphic representation of the model reconstruction and the structural recognition of Fig. 7.a
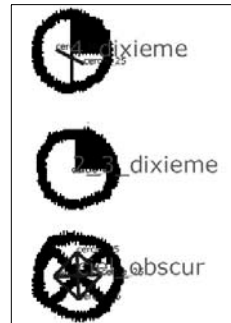
The matching graph tool used does not allow the localization and therefore the manipulation of sub-graphs in a candidate graph. Thus, it is impossible to exploit a candidate graph representing the whole image loops. For that reason, it is impossible to distinguish the symbols "*chambre*" when they are connected to the symbols "*PC paires sur appui*" and "*PC paires sur bornes*". Indeed the "*chambre*" are closely connected to this PC (Fig. 5.a and Fig. 2.a), and a too strict distance constraint could split up the PC symbol in several symbols. To overcome this problem, we have considered the symbol "*PC paires sur support + chambre*" and "*PC paires sur appui + chambre*" as complete symbols. We didn't have this problem with meteorological symbol because distance between loops is sufficiently long.

3.5 Strategy and Results

In the setting of our application, we have tested several symbol recognition strategies. The most efficient strategy uses the models only taking into account the distance constraint. Indeed, recognition errors in statistical classification inevitably generate some errors in the models reconstruction, if one takes into account connection constraints. With this strategy, we obtained completely inter-connected graph for every symbol. The matching tool is then parameterized to only take into account the global similarity criterion between nodes (Fig. 7.b). This global criterion is a weighted average of two similarity criteria, the first on topology graphs, and the second on exact graphs. The weighted average is computed with a more important coefficient for similarity criterion on graphs topology.

In term of results, we have tested this approach on 29 symbols, constituted of about hundred loops, and distributed on 9 plan extracts. We have constituted a training set of feature vectors describing the loops. This training set has been used as test set, in order to obtain a 100% statistical recognition results. We obtained 100% of recognition on symbols.

Obviously, we are interested to test the correction ability of structural recognition step on statistical recognition results. We voluntary altered statistical training set of France Telecom symbols loops in order to reduce the loops recognition rate. Tests carried out on 22 symbols composed of 74 loops give a loops statistical recognition rate of 55.4% and a symbol structural recognition rate of 86.86%. Among these symbols we have up to 75% of a wrong statistical recognition loops. These results prove the ability of structural recognition step to correct the statistical recognition step. But, taking into account similarity criterion on topology graphs increases these results. Indeed, only the node number used alone permits to distinguish France Telecom symbols (1 for "*chambre*", 2 for the "*PC paires sur appui*", 3 for the "*PC paires sur appui + chambre*", 4 for the "*PC paires sur bornes*", 5 for the "*PC paires sur bornes + chambre*"). We realized similar tests on the meteorological symbols in order to prove the importance of similarity criterion on topology graphs in symbol recognition. Tests on 56 symbols composed of 96 loops extracted from the same image rotated into 8 different directions gives 44% for loops statistical recognition and 55% symbol structural recognition. Indeed, only symbol "*ciel obscurci*" (Fig. 2) can be recognized by its loops number. Nevertheless, we have corrected 33% of symbol in which loops were badly recognized by statistical step.

3.6 XML Modelling

Data manipulated by our system are modelled in XML [15]. The use of this data de-scription language offers several advantages. First, XML seems to become a refer-ence concerning data description languages. This guarantees in the future a durability of our tools and especially a possibility to exploit tools provided by the scientific community. Secondly, the properties of XML can be used in a recognition system. Among these properties, we used the data transformation and the specialized sub-languages.

Data transformation tools: XML permits to use of XSLT processors (for example, the Xalan processor [16]). These processors transform a XML data flow with the help of a XSLT script program [17], which permits an easy data transformation. If data have a XML type flag format, we can have two tools using different communication data structures. In the same way, it is also possible to merge data stemming from several tools.

The specialized sub-languages: XML is described as a meta-language because it is a root language, which permits to define specialized sub-languages. For example the SVG [18] language permits a data graphic description. We used this language in order to rebuild a graphic representation of all our steps processing (Fig. 5.b, 6). Moreover, we can superpose the image with the graphic representation of our process result (Fig. 7.b). With the help of the XSLT processor, we merged and transformed information of our different tools in SVG format. We used tools provided by com-puter science community (SVG viewer [19], Batik [20]) for the recognition visualiza-tion.

4   Conclusions and Perspectives

In conclusion, we have presented in this document a symbol recognition system com-bining statistical and structural approaches. We have exploited these approaches in order to recognize some technical symbols composed of loops. These symbols have proximity relations between theirs loops. We exploited statistical approach in a first interpretation level in order to recognize loops found in symbols. In a second interpretation level, we exploited proximity relations between these loops by a structural approach, in order to recognize the complete symbols. The first results are encouraging. On one hand, a perfect statistical recognition gives a perfect symbol structural recognition. On the other hand, the statistical recognition results can be corrected by structural recognition step. The efficiency of this correction is a function of similarity between symbols (differences in topology symbols, share of loops classes between symbols). The model set and the configuration information of tools have been outsourced from algorithms. This confers to the system a generic aspect. Data manipulated by our system are modelled in XML. This gives to system the ability to interface tools using different data format, and to create graphic representation of each treatment step.

Our first perspective is to extend the statistical/structural serial combination in a par-allel combination. Indeed, we hope will integrate structural model extraction tools for compare with statistical model extraction tolls actually used. These structural tools

allow extracting connected component structure (contrary to statistical model extraction tolls). This property gives ability to extract part of a connected component. Among these structural model extraction tools we hope will integrate:

- x squeletonization method - skeleton structuring - and mathematics approximations [21]
- x follow line method - structuring method of follow line [22]
- x line adjacency graph method [23]

The second perspective consists in improving our structural classifier. In a first time we plan to realize localization and manipulation of sub-graphs in a candidate graph, in order to exploit loops graph representing several symbols on image. It will permit to treat the neighbor symbols that can't be isolated by simple distance constraint (as it's the case of the France Télécom symbol "*chambre*"). In a second time we hope will compute inexact graph matching in order to allow a tolerance between node and edge values of candidate and model graphs during matching process. It would permit for example to take into account the distances between loops given by reconstruction models tool. Finally, in a third time, we wish to exploit the confusion matrix of our statistical classifier. The goal is to weight similarity computation between graphs according to confidence degrees of node labels (provided by the statistical classifier).

Finally, the third perspective is to integrate our tools in a knowledge-based approach. In a first time, the aim is to create a common knowledge set in XML, for the whole tools of our system. A simultaneous use of XSLT and XML-QL [24] (norm for the XML data set management) will permit management and adaptation of this knowledge set for the whole tools of our system. It will be necessary to define a representation formalism of a generic model for the entire recognition tools. In a second time we wish to develop a supervision program of our recognition system, permitting the combination of our different tools (classifiers and models extraction tools). The aim is to control the whole processing from the image processing to the classifier combination. We wish to control our recognition system by a processing scenario, in order to adapt easily and quickly the system to a new recognition objective.

References

1. L. Boatto and al, An interpretation system for land register maps, IEEE Computer Magazine, 25(7), pp 25-33, 1992.
2. S.H. Joseph, P. Pridmore, Knowledge-directed interpretation of line drawing images, IEEE Trans. on PAMI, 14(9), pp 928-940, 1992.
3. J.M. Ogier, R. Mullot, J. Labiche and Y. Lecourtier, Multilevel approach and distributed consistency for technical map interpretation: application to cadastral maps, Computer Vision and Image Understanding (CVIU), 70, pp 438-451, 1998.
4. P. Vaxivière, K. Tombre, CELESTIN : CAD conversion of Mechanical Drawings, IEEE Computer Magazine, 25, pp 46-54, 1992.

5. K. Chhabra, Graphic Symbol Recognition: An Overview, Lecture Notes in Computer Science, vol. 1389, pp 68-79, 1998.
6. J.Lladós, E. Valveny, G. Sánchez, E.Martí, Symbol recognition: current advances an perspectives, 4th IAPR International Workshop on Graphics Recognition (GREC'01), Kingston, Canada, 1:109128, 2001.
7. B. Pasternak, B. Neumann, Adaptable drawing interpretation using object oriented and constrained-based graphic specification, in proc. Second International Conference on Document Analysis and Recognition, Tsukuba, Japan, pp 359-364, 1995.
8. N.A. Langrana, Y. Chen, A.K. Das, Feature identification from vectorized Mechanical drawings, Computer Vision and Image Understanding, 68(2), pp 127-145, 1997.
9. G. Myers, P. Mulgaonkar, C. Chen, J. Decurting, E. Chen, Verification-based approach for automated text and feature extraction from raster-scanned maps, in Proc. of IAPR International Workshop on Graphics Recognition, Penn State Scanticon, USA, pp 90-99, 1995.
10. S.W. Lee, Recognized hand-drawn electrical circuit symbols with attributed graph matching, in H.S. Baird, H. Bunke, K. Yamamoto, eds., Structured Document Analysis, Springer Verlag, pp 340-358, 1992.
11. B. Messmer, H. Bunke, Automatic learning and recognition of graphical symbols in engineering drawing, in R. Katsuri and K tombre, eds., Lecture Notes In Computer Science, volume 1072, pp 123-134, 1996.
12. R. Schettini, A general purpose procedure for complex graphic Symbols Recognition, Cybernetic and System, 27, pp 353-365, 1996.
13. S. Adam, J.M. Ogier, C. Cariou, J. Gardes, Y. Lecourtier, Combination of invariant pattern recognition primitive on technical documents, Graphic Recognition – Recent Advances, A.K. Chabbra D. Dori eds., Lecture notes in Computer Science, Springer Verlag, vol 1941, pp 29-36, 2000.
14. P. Héroux, S Diana, E. Trupin, Y. Lecourtier, A structural classification for retrospective conversion of document, Lecture Notes in Computer Sciences, Springer Verlag, vol 1876, pp 154-162, 2000.
15. World Wide Web Consortium, eXtensible Markup Language (XML) 1.0, http://www.w3.org/TR/2000/REC-xml-20001006 , 2000.
16. Apache XML projects, Xalan processor 2.2 D14, http://xml.apache.org/xalan-j/index.html
17. World Wide Web Consortium, eXtensible Style-sheet Language Transformation (XSLT) 1.0, http://www.w3.org/TR/xslt , 1999.
18. World Wide Web Consortium, Scalable Vector Graphic (SVG) 1.0, http://www.w3.org/TR/SVG/ , 2001.
19. Adobe, Svg Viewer 3.0, http://www.adobe.com/svg/
20. Apache XML projects, Batik SVG toolkit 1.1, http://xml.apache.org/batik/
21. X. Hillaire, K. Tombre, improving the accuracy of skeleton-based vectorisation, IAPR International Workshop on Graphic Recognition (GREC), Kingston, Canada, 2001.
22. J.M. Ogier, C. Olivier, Y. Lecourtier, Extraction of roads from digitized maps, in processing of the sixth EUSIPCO (European Signal Processing Conference), Brussels, Belgium, pp 619-623,1992.
23. S. Di Zenzo, L. Cinque, S. Leviadi, Run-based algorithms for binary image analysis and processing, IEEE Trans. on PAMI, 18(1): 83-89, p56, 1996.
24. World Wide Web Consortium, XQuery 1.0 an XML query language, http://www.w3.org/TR/xquery/ , 2001.