RIST

Information Processing at the Digital Age

Vol22.N.2(2018) 61-65

A Neural Model for content-based XML Information Retrieval

F.Z. Bessai-Mechmache*, T. Abdi, A. Hadibi

Research Centre on Scientific and Technical Information, CERIST, Ben Aknoun, 16306, Algiers, Algeria

Submitted 27/01/2018, accepted 20/03/2019

Abstract

The aim of a content-based XML information retrieval system is to select relevant XML elements according to user query expressed by keywords. The main issue for this system is how to select relevant XML elements belonging to various XML document sources in relation to a given query. To do this, we propose a neural XML information retrieval model using Kohonen self-organizing maps. Kohonen self-organizing map produces density map that form the foundations of the XML information retrieval system.

Keywords: Neural Networks; Self-organizing maps; heterogeneous XML data; XML Information Retrieval.

1. Introduction

The strategic interest carried in the information as well as the explosive advent of the Internet and other information services are key factors supporting the growth of research directions aimed at implementing automatic process access to information constantly more efficient. Also, the development of electronic document and Web have emerged then impose structured data formats such as XML (eXtensible Markup Language), to represent information in a richer than the simple form and content adapted to specific needs. These new formats allow representing jointly the textual information and the information of structure of a document. The Knowledge of the structure of documents is an additional resource that should be exploited for information retrieval in order to better answer a need for information (Kamps, 2003; Fuhr, 2004).

The logical structure or hierarchy of XML documents contains content and structural elements. The purpose of an XML information retrieval system is refined to retrieval strategies, which aim at returning document components, i.e. XML elements, instead of whole documents in response to a user query (Sigurbjornsson, 2003).

* F.Z. Bessai-Mechmache. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: zbessai@cerist.dz

Content-only queries make use of content constraints only, i.e. they are made of keywords and are suitable for XML retrieval scenarios where users do not know, or are not concerned, with the document logical structure when expressing their information needs. In this paper, we define a content-based XML information retrieval model, which cope with heterogeneous XML data. To do this, we suggest an XML information retrieval model, based on the classification of structural elements of XML documents using Kohonen Self-Organizing Maps (SOM).

This paper is structured as follows. We initially introduce related works concerning XML information retrieval. We describe the proposed model in section 3. In section 4 we conclude the paper.

2. Related Works

All Many works in literature deal with XML information retrieval. Approaches suggested addressing the problem of relevant elements selection when queries contain only keywords can be divided into two main categories (Abolhassani, 2004). These approaches are characterized by the way they index the elements of an XML document.

The first category considers that any element (or sub-tree of elements) can potentially be returned to the user. Each element is thus treated like an atomic unit. The smallest sub-trees are those that only include leaf of tree. In order to identify the elements to be returned in response to the query, a score of relevance is calculated between the query and each sub-trees. The sub-trees results then are sorted and returned to the user according to their scores of relevance (Abolhassani, 2004; Sigurbjornsson, 2003).

The second category considers that elements to be returned to the user are not necessarily leaf nodes. Each element of a document can be returned to the user, but the relevance score of a given element with respect to a query is computed by aggregating the scores of relevance of its child elements. Several models were proposed within this framework (Agrawal, 2009; Arguello, 2011; Bessai-Mechmache, 2011; Bessai-Mechmache, 2012, Clarke, 2008; Govert, 2002; Lalmas, 2012; Ogilvie, 2003; Sauvagnat, 2006].

However, the approaches mentioned above deal with XML documents having same hierarchical structure (same XML schema). In this paper we will expand the coverage to take into account both homogeneous and heterogeneous XML documents while focusing on how to select relevant XML elements according to user query expressed by keywords. For this purpose, we propose an XML information retrieval model based on neural networks (Haykin, 1999; Oevermann, 2018) and more particularly on Kohonen self-organizing maps (Kohonen, 1990; Kohonen, 1995). These maps allow a thematic clustering of XML elements producing density map that form the foundations of XML search.

The main feature of the proposed model is the schema heterogeneity where documents can have different hierarchical structures (different XML schemas). To deal with heterogeneity the selection process instead of document handles parts of document (XML elements) which makes the number of information to be processed very important. One approach to cope with this challenge is to use clustering algorithms, which can contribute better performing XML information retrieval systems (Suma, 2014). Kohonen's self-organizing map (SOM) (Haykin, 1999; Kohonen, 1990; Kohonen, 1995) is one of the most well known methods that can help to achieve this goal. The idea is to apply the self-organization phenomenon of XML elements all around the user query. For this purpose, we suggest a selection model using Kohonen self-organizing maps. These maps gather XML elements in clusters dealing with same topic and thus reduce the amount of data to be process to select relevant information.

3. Neural XML Information Retrieval Model

3.1 Model Architecture

The architecture of the proposed model is a Kohonen neural network that consists of a grid (map) of neurons and an input stimulus, as illustrated in Figure 1 below.

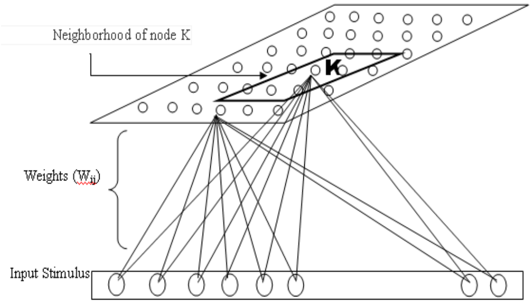

Fig. 1. Model architecture

3.1.1 Map Description

Each neuron of the map depicts an element of an XML document. The map is divided into subset of XML elements (areas or clusters) having the same characteristics, i.e. dealing with the same topic. Therefore, elements belonging to the same cluster are potential candidates to appear in the same result list for a given query.

Input stimulus is an N-dimensional vector that represents an element to be classified on the map.

The input stimulus consists of indexing terms of a given element.

Connection that join every term of the stimulus to the map reflect the importance (W_{ij}) of term 't_i' in element 'e_j'. This importance (or weight) is calculated by following equation (Trotman, 2005).

$$W_{ij} = tf_{ij} * ief * idf \qquad (1)$$

With 'tf_{ij}' term frequency, 'ief' inverse frequency of element 'e_j' for term 't_i' and 'idf' inverse frequency.

3.1.2 Learning Algorithm

Let $X(t) = \{ X_1(t), X2(t),\ldots, Xn(t)\}$ be a learning pattern at instant t.

Let $W^k(t) = \{ W_1^k (t), W_2^k(t) ,\ldots, W_n^k(t)\}$ be a neuron at instant t.

Firstly, the map must be initialized randomly.

For each input pattern:

1- Calculate the distance between the pattern and all neurones ($\|X(t)-W^k(t)\|$). The chosen distance measure is the Euclidean distance.

2- Select the nearest neurone as winner W^s ($\|X(t)-W^s(t)\| = \min \|X(t)-W^k(t)\|$)

3- Update each neurone according to the rule:

$W^k_i(t+1) = W^k_i(t) + \alpha(t).h_{(W^s, W^k)}(t).\|X_i(t) - W^k_i(t)\|$ with $1 \leq i \leq n$.

Let $0 \leq \alpha \leq 1$ be the learning rate, and $h_{(W^s, W^k)}$ be the neighbourhood function. This function assumes values in [0, 1] and is high for neurones that are close in the neighbourhood, and small (or 0) for neurones far away.

4- Repeat the process until a certain stopping criterion is met. Usually, the stopping criterion is a fixed number of iterations.

To guarantee convergence and stability of the map, for each iteration the learning rate and neighbourhood radius are decreased, thus converging to zero.

3.2. Query processing

A query consists of a set of terms, $Q = (t_1, t_2, ..., t_m)$.
In our approach, we consider the query as a vector of terms ready to be classified on Kohonen map.

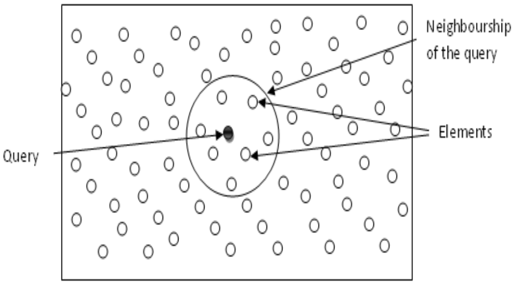

Fig. 2. Classification of the query on Kohonen map

Once the query classified on the map, the XML elements belonging to the cluster of the query are used to compute the relevance of XML elements that answer the user's query.

The relevance of an XML element 'e', with respect to the query 'Q' is equal to the sum of the weights of query terms with regard to this element. This relevance is denoted RSV(Q, e) and is calculated as follows:

$$RSV(Q, e) = \sum_{i=1}^{M} W_i^e \tag{2}$$

With:
- W_i^e : weight of term 't_i' in element 'e'.
- M: number of query terms.

Once relevance of all XML elements belonging to the cluster of the query calculated, the result returned to the user is a list of elements ordered according to their degrees of relevance

4. Conclusion

In this paper, we proposed a formal framework for a content-based XML information retrieval system, which copes with heterogeneous XML data. The proposed model thanks to Kohonen self-organizing map selects XML elements from different parts of various XML documents to build a ranked list of relevant XML elements. This relevance is based on the content and structure of XML documents.

Future work will concern the evaluation of the proposed model on a data set.

References

Abolhassani, M. and Fuhr, N., 2004. Applying the divergence from randomness approach for content-only search in xml documents. In Proceedings of ECIR 2004, Sunderland, pages 409–419.

Agrawal, R., Gollapudi, S., Halverson, A., 2009. *Diversifying Search Results*, ACM Int. Conference on WSDM.

Arguello, J., Diaz, F., Callan, J., 2011. *Learning to aggregate vertical results into web search results.* In Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM'11, Glasgow, United Kingdom.

Bessai-Mechmache F.Z., Alimazighi Z., 2011. Possibilistic Networks for Aggregated Search in XML Documents, in proceedings of International Conference on Information & Communication Systems, ICICS'2011, p.67-72, ISBN 978-1-4507-8208-1, Irbid, Jordan.

Bessai Mechmache, F.Z., Alimazighi, Z., 2012. Possibilistic model for aggregated search in XML documents. IJIIDS 6(4): 381-404.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., 2008. *Novelty and diversity in information retrieval evaluation*. SIGIR'08, p.659-666.

Govert, N., Abolhassani, M., Fuhr, N. and Grossjohann, K., 2002. Content oriented XML retrieval with hyrex. In Proceedings of the first INEX Workshop, Dagstuhl, Germany.

Haykin, S., 1999. *Neural Networks: A Comprehensive Foundation*, Prentice Hall, ISBN 0-13-273350-1.

Kamps, J., Marx, M., De Rijke, M., Sigurbjörnsson, B., 2003. *XML Retrieval: What to retrieve?* ACM SIGIR Conference on Research and Development in Information Retrieval, p.409-410.

Kohonen, T., 1990. *Self-organizing map*. Proceedings of the IEEE, Vol. 78, n°9.

Kohonen, T., 1995. Self-organizing map. Volume 30 of Springer Series in Information Sciences. Springer, Berlin, Heidelberg.

Lalmas, M.. 2012. Xml information retrieval. Understanding Information Retrieval Systems: Management, Types, and Standards.

Fuhr, N., Lalmas, M., Malik, S., Szlavik, Z., 2004. *Advances in XML Information Retrieval: INEX 2004*. Dagstuhl Castle, Germany, December 6-8.

Oevermann, J., Ziegler, W., 2018. Automated classification of content components in technical communication. Computational Intelligence, 34:30–48.

Ogilvie, P., Callan, J., 2003. *Using language models for flat text queries in XML retrieval*. In Proceedings of INEX 2003 Workshop, Dagstuhl, Germany, p.12-18.

Sauvagnat, K., Boughanem, M., Chrisment,, C., 2006. *Answering content-and-structure-based queries on XML documents using relevance propagation.* Information Systems, Special Issue SPIRE 2004, vol. 31, p.621-635, Elsevier.

Sigurbjornsson, B., Kamps, J., de Rijke, M., 2003. *An element-based approach to XML retrieval*. INEX 2003 workshop, Dagstuhl, Germany.

Suma, D., U. Dinesh Acharya and GeethaM , Raviraja Holla M., 2014. XML Information Retrieval: An overview, International Global Journal for Engineering Research, Volume 10 Issue 1.

Trotman, A., 2005. *Choosing document structure weights*. Information Processing and Management, vol. 41, n°2, p.243-264.