Automatic Hypertext Construction in Persian Texts Using Self-Organizing Map Neural Network

Mahdieh HajiMohammadHosseini, Department of information technology, university of Qom, Qom, Iran
m.hajihoseini@stu.qom.ac.ir
Behrouz Minaei-Bidgoli, Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
b_minaei@iust.ac.ir

Paper Reference Number:07-01-2000
Name of the Presenter: Mahdieh HajiMohammadHosseini

Abstract

With the availability of electronic texts, users are encouraged to study them. Therefore users may encounter during their study with different information needs and want more information or related information about a particular word or phrase within that document. If so, it is necessary to search the entire corpus of texts and then they are faced with problems related to the search. Using hypertext is a fast method for retrieving information. Manually converting large amounts of documents into hypertext is time consuming and sometimes impossible. The purpose of this paper is to implement an automated way to convert texts into hypertext. This is the first activity and implementation in Persian documents. In this approach, two types of links are made using Self-organizing Map neural network, two labeling processes and analyzing them. In this study, in addition to single words links, two-word phrases links are produced too. Some of links sources in generated links were in title of the destination document that shows high correlation between source and destination; but about other sources, we specified most related paragraph in the destination document. The average precision rate of the two types of links for single words and phrases was calculated 0.71.

Key words: Automatic hypertext construction, Information Retrieval, Neural networks, Text mining

1. Introduction

With the increasing use of computers and the Internet, the amount of available electronic information has increased. Electronic information on the Internet is in various forms like text, audio, image, video and etc. However, the most numerous and most widely used form of electronic data is text. Hypertext is a text of words segments (or images) that connected electronically by multiple paths, chains or sequences in an infinite and interconnected environment.

The main challenges in this issue are: (1) Finding useful words or phrases within the text for link source and (2) Finding an article or part of it that the link points to. Link discovery approaches based on applied knowledge are divided in three categories: link based methods, title based methods and concept based methods.

In this project with emphasizing a method based on text content, we want to implement an automatic hypertext construction method with acceptable performance. In this study, the self-organizing map (SOM) algorithm is used for documents clustering. Through two labeling processes, documents associated with each neuron and important words of documents in same neurons are indicated.

Since human doesn't interfere in the activities, a lot of errors related to human intervention are avoided. Another advantage is the ability to insert and delete documents in the automatic operation, while in manual construction of hypertext, insert and delete a document by the author is very time consuming and even impossible processes. In the automatic creation of hypertext, the consistency and stability of links is more than time with human intervention. Automatic conversion of electronic texts into hypertext is a quick process.

Due to the benefits of using hypertext, automatic production and as well as problems related to manual construction of hypertext, efforts have been made towards automatic conversion of text to hypertext.

Salton and Buckley (1989) used a content based method. They looked for similar pairs of documents and then among the most similar cases, most similar sentences were linked. In Salton and Singhal (1995) research, the queries are compared with each document vector by inner product and most similar document is selected. Agosti et al. (1996) designed TACHIR model for this purpose. This model includes document, index words and concept layers. Through these layers and links created among them, hypertext is constructed. Crestani and Melucci (2003) implemented TACHIR practically. Green (1998) presented a model based on lexical chains for solving problems related to links of synonyms words.

Zeng and Bloniarz (2004) used keywords for generating links. Links were generated based on similarity of paragraphs contain keywords. Lee and Yang (2005) designed a model based on using SOM neural network. In this method two types of links are created: links from index words of documents to similar documents and links from less important words to documents describing that word. Using machine learning technic, Milne and Witten (2008) generated new links and disambiguated Wikipedia links. In this study, C4.5 algorithm has been used.

Granitzer, Seifert and Zechner (2009) selected titles as link candidates. Zhang and Kamps (2009) were looking for creating links in Wikipedia according to duplicate links. Link generation in Xiong and Gardner (2009) project was based on labeling words and using classifier. Knoth et al. (2010) generated links between text segments by calculating semantic similarity. Other project by Geva, Trotman and Tang (2010), were applied combining ICLM and GPNM algorithms for finding links on Wikipedia.

Alarabi (2011) uses information retrieval measure and neural network for calculating similarity. Itakura et al. (2011) first generated links according to words frequency and then applied KPR on titles to create new links; second process show better results.

Among the activities carried out for text mining, Self-organizing maps (SOM) for clustering the documents is used a lot. For example, the following can be noted: Kaski and et al. (1998), Lee and Yang (1999), Rauber and Merkl (1999). Also Isa et al. (2009) used Naïve Bayes for converting documents to vector by probability distribution and then applied SOM for unsupervised classification. Other new activities in using SOM for text organizing include projects of Liu, Wang and Wu (2008), Chandrashekar and shoba (2009), Matharage et al. (2011) and Yang and Lee (2012).

Nevertheless, so far, such activities for automatic construction of hypertext did not occur at Persian documents. This study tried to create a solution to automatically generate links based on text content in the Persian texts.

The structure of this paper is as follows. We will describe data used in this study in section 2. In section 3 explain processes of hypertext construction. In section 4, we express results from implementing method. Finally we present conclusions in section 5.

2. Data and Material

The data set used in this study provided from the articles of Research Center of Islamic Sciences. This collection includes 1988 articles that some preprocessing tasks were performed before using them. As mentioned previously, so far, Persian documents were not researched in the field of converting plain texts to the hypertext. In this study, we applied the approach used in Lee and Yang (2005) in order to construct hypertext in Persian documents. Finally we determined most related paragraph in the destination documents.

This activity was performed in three steps:

- Documents preprocessing
- Clustering documents and labeling clusters
- Link generation

3. Research Methodology

Text mining processes was done in some steps as follows: in Section3-1 we express documents preprocessing and steps for converting documents to vectors. We explain clustering by using SOM in Section3-2. Section3-3 contains contents about cluster labeling. Finally Section3-4 illustrates two types of links generation.

3-1. Preprocessing

The purpose of the pre-processing is converting text from irregular and unstructured form to regular and structured form. Since in this study, there were Arabic words in documents text, first, the entire text was broken to its sentences. Then every sentence was passed to text language detection program, because the aim of the project is Persian sentences processing solely; otherwise the number of index words in a document i.e. the dimension of feature vector, increase and consequently complexity increase.

One of the important pre-processing steps is part-of-speech tagging. We want to choose nouns as link source candidates; because it is believed nouns carry most of semantic in the text. Then plural words must be converted to singular mode. Plural nouns with suffixes, "**ان**", "**ها**", "**ات**", "**ين**", transformed to singular mode. Next, stop words, frequent and infrequent words were removed.

Index words were weighted by TF-IDF method. This value was calculated as Eq. (1)

$$Tfidf(t,d,D) = tf(t,d) \times idf(t,d) \quad (1)$$

tf(t,d) represents frequency of word t in document d and idf(t,d) means inverse document frequency. Vector space model is a model that has many applications in information retrieval and text mining. In this model, each document or a piece of text is represented as a vector of words such that each member of this vector shows the weight of each word in the document.

3-2. Clustering

SOM is an artificial neural network that follows unsupervised learning. This algorithm is useful for visualization and interpretation of data with high dimensions and mapping them to a lower dimensional space. In this neural network, After training, data are placed inside neurons such that similar data (i.e. documents) are placed in same neuron. On the other hand, in SOM, neighboring neurons are closely related. Neurons or clusters are connected by a neighborhood relationship such that this relationship defines the topology or structure of the map. In this article, sometimes we call neurons as clusters.

If $X_i=\{x_{in}|1<n<N\}$, $1<i<M$ represent inputs of neural networks such that N is the number of index words or in other words the dimension of document vector. M determines documents count. SOM will form a regular grid of neurons that synaptic weight vector of neurons defined as $W_j= \{w_{jn}|1<n<N\}$, $1<j<J$. each neuron has N synapses, i.e. the number of index words in each document vector, that represent weight of corresponding index word in that neuron. J is the number of neurons in the network. At the beginning of the learning process, the weight vectors are randomly initialized. Network learning is done in the loop below:

1- Randomly select one of the vectors of the input documents i.e. x_i. Selected vector should not be chosen before in same epoch.
2- Find a neuron such that its synaptic weight vector has minimum distance with selected document vector, i.e. x_i. For finding this neuron, the Eq. (2) must be true:

$$\| x_i- w_j\|= \min \| x_i- w_j\|; 1<k<J \quad (2)$$

$\| x_i- w_j\|$ is calculated as Eq. (3):

$$\| x_i - w_j \|= \sqrt{\sum_{i=1}^{N}(x_i - w_j)^2} \quad (3)$$

3- Update weight vectors of neighborhood neurons of selected neuron, as shown in Eq. (4).:

$$w_t^{new} = w_t^{old} + \alpha(t)(x_i - w_t^{old}) \quad (4)$$

$\alpha(t)$ is training gain in epoch t.
4- Repeat steps 1-3 until all vector documents are chosen.
5- Increase t. If the number of iterations exceeds a preset value, the operation ends. Otherwise decrease $\alpha(t)$ and neighborhood size and go to step 1.

3-3. Labeling Processes and Clusters Analysis

After training process, two labeling processes are done on clusters. Labeling helps us in link generation. First labeling process is document cluster map (DCM). DCM is mapping documents to clusters. For each document, Eq. (1) is checked in order to find closest synaptic weight vector. After identifying closest to with a document, the document labels to closest neuron. Due to high common words, documents with similar semantic themes place in same neuron or neighborhood neurons. On the other hand if the number of neurons is too high compared to the number of documents or documents theme is relatively similar, it is possible some neurons are not labeled.

Word cluster map (WCM) is second labeling process on clusters. It is assumed that when a cluster learns a word well and knows its importance, it has higher weight for that word in its synaptic weight vector. Thus by specifying a threshold for weight, we can determine which neurons must be labeled to a cluster.

Because in this study, training documents had similar theme mostly and documents distribution in neurons was variant, weight vector of some neurons had high values and some neurons had lower values in their weight vectors. Thus identifying threshold for labeling most neurons, was caused some neurons that have weight vector's average higher than threshold are labeled with too words. On other hand some neurons are not labeled or are labeled with few words. So in this study, instead of identifying threshold, we specify the number of words for labeling to neurons. We consider more words for labeling, for clusters with higher weight average, because we believe these neurons learn more words and can be good guidance for link generation.

Finally unlabeled words assign to a neuron with maximum weight for that word. In learning process, words are clustered based on words co-occurrence. When two certain

words repeat mostly together, learning process try to learn these words simultaneously (e.g. words "**ترجمه**" and "**قرآن**"). Therefore some words repeat simultaneously is labeled to same neurons or neighboring neurons; of course, the opposite is also true. Since documents labeled to same neuron have minimum distance with weight vector of neuron, then these documents vectors are normally close together and share some words. As a result, words labeled to same neuron shows general concept of documents labeled to same neuron. These words set can be identified as a thesaurus.

3-5. Link Generation

Links are generated by analyzing labeling processes. Two main steps in link generation are: (1) finding link source by WCM analyzing and (2) specifying link destination by DCM analyzing.

In this activity two types of links are constructed: inter cluster links and intra cluster links. Sources of inter cluster links are not keywords or index words of a document. Readers often tend to follow these links, because they want to know more concepts about words that document do not explain them enough. Often these links point to a document with a different theme to the source document such that the link source is an index word in it and somehow express the subject of the destination document. With this explanation, it is reasonable that the destination document is not in same neuron with source document, because documents in same neuron have relatively similar semantic.

Intra cluster link start from index words or keywords of documents. The source is descriptive of document theme. The source of link is an important word in a document that the intra cluster link point to it too. Since index words of two documents are common, documents have relatively similar themes. We expect that source and destination of an intra cluster link have same cluster because they have similar subjects.

3-5-1. Finding Source

First, we explain finding source of inter cluster link. As mentioned before, source of these links is not index word of the document. Index words or important words of documents labeled to document cluster. For creating link for document D_j in neuron c with labeled words W_c; source k_i must have following conditions in Eq. (5):

$$k_i \notin W_c, k_i \in W_m, m \neq c \tag{5}$$

W_m is words set labeled to neuron m.

For sources of intra cluster links, we should select word k_i in document D_j as Eq. (6). D_j labeld to neuron c with labeled word set W_c.

$$k_i \in W_c \tag{6}$$

To avoid excessive links production, we defined σ_1 and σ_2 to limit the number of links. Words or phrases that have been identified as sources of inter cluster links, are ranked according to Eq. (7) and then σ_1 top ranked source candidates are selected.

$$\overline{w}_{ic} = \frac{1}{|C_i|} \sum_{m \in c_i} w_{im} \tag{7}$$

In this equation, C_i represents clusters that labeled with k_i. For intra cluster links, it is enough to select σ_2 words that have highest component values in the synaptic weight vector of the neuron corresponding to this word cluster.

3-5-2. Finding Destination

Below requirements are necessary for finding destinations of inter cluster links with source k_i in document D_j.

1- D_j and D_{ki} should be belonging to different clusters; D_{ki} represent destination of source k_i. As previously mentioned, different clusters are reasonable for this type of link.

2- $k_i \in W_m$, $m \neq c$ and

$$w_{im} = \max w_{il}; l \neq c, 1 \leq l \leq M \quad (8)$$

c is index of the neuron that contains D_j. This requirement is for finding most related cluster with k_i. Because we want to find maximum weight corresponding k_i in all cluster vectors.

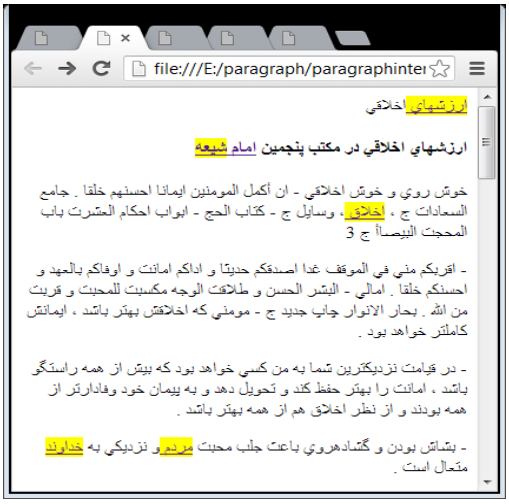

Fig. 1. An example of single words links

Fig. 2. An example of phrases links

3- The distance between D_j and D_{ki} should be minimum value in order to select most similar document with source document, as Eq. (9).

$$\| x_{kj} - x_j \| = \min \| x_l - x_j \|; 1<l<M \quad (9)$$

But for finding destination of intra cluster link, since we are looking for a document close to source document, we search the same source document cluster for finding link destination. For this purpose we select a document from same cluster that it has maximum weight for k_i between all documents in that cluster.

4. Results and Analysis

In this study, after preprocessing, number of distinct words became 14862 words. Here we extracted two-word phrases by considering co-occurrence of words. The process of hypertext construction was performed on phrases too. Articles employed, were in Persian language. First paragraph of each document contains its title. Each document was transformed to a vector with 14286 elements. The method of word weighting was TF-IDF. Since most documents were on a semantic domain, the use of large numbers of neurons, leading to a large number of neurons were not labeled with any document. For this reason, we applied a 6*6 grid for training so that there are no unlabeled neurons. We set the maximum epoch to 100 and initial training gain to 0.0001.

After training, neurons were labeled with DCM and WCM. As expected, because of similar theme in many documents, some neurons labeled with more documents. Association of WCM tag with subject of documents within the cluster was observed well.

Because the average length of documents was large, the maximum number of links was set to 30 and 20 for intra and inter cluster link respectively.

Fig. 1 shows an example of inter cluster links of single words. Fig. 2 indicates a sample for intra cluster links of phrases. As can be seen in the figures, some links have a colored background that means existence of this word in title of destination document and probably source document has acceptable relationship with destination document. Although it is possible if the source word does not exist in title of target document, related to it and link was created truly. Since documents had high length and sometimes they contained different contents; we specified most related paragraph with source of link. This

paragraph chooses based on frequency of link source. The paragraph was selected with maximum frequency of source. Determining this paragraph is more useful about sources that do not exist in the title of destinations. Because readers wish to access a text segment related to link source.

Finally we wanted 5 users to evaluate samples of results. They were asked if there are beneficial relationship between the source and destination of the link, evaluate it as true and otherwise false. Finally we calculated precision measure for evaluation results of each user. Eq. (10) shows the formula of precision:

$$precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{10}$$

Then we calculated average of results. Results are shown in Table 1.

This project was first implementation on Persian documents that approved by expert and the precision achieved is acceptable. Since data were not linked before, we cannot calculate recall measure.

Link type	Precision
Single word intra cluster	0.66
Single word inter cluster	0.78
Phrase inter cluster	0.73
Phrase intra cluster	0.7

Table 1. The precision of generated links

5. Conclusions

In this activity, for first time, we wanted to implement a method of converting plain texts to hypertext using SOM neural network. The purpose of implementation was generating two types of links; intra cluster link and inter cluster link. These links were produced by analyzing DCM and WCM labeling process. Intra cluster links connect source document to a document that has similar theme. These links start with important and index words. In contrast, inter cluster links originate from less important words of documents and are connected to a document that contains more detail about source word.

The implementation was performed for single words and also two-word phrases. Precision achieved is acceptable and approved by expert. In future work, since structure of Persian language has high complexity more studying and preprocessing can be done to decrease the number of index words. Also better and more useful words set are produced by implementing a method for stemming nouns.

References

Agosti, M. et al. (1996). Design and implementation of a tool for the automatic construction of hypertexts for information retrieval. *Information Processing & Management*, 32(4), 459-476.

Alarabi, A. (2011). Building Automatic Hypertext Links Using Artificial Neural Models. *Journal of Information Organization*. 1(1), 17-22.

ChandraShekar, B.H. and Shoba G. (2009). Classification of documents using kohonen ' s self-organizing map. *International Journal of Computer Theory and Engineering*. 1(5), 610-613.

Crestani, F. and Melucci, M. (2003). Automatic construction of hypertexts for self-referencing: the Hyper-TextBook project. *Information Systems*, 28(7), 769-790.

Erbs, N. et al. (2011). Link discovery: A comprehensive analysis. *In Proceedings of the 5th IEEE International Conference on Semantic Computing (IEEE-ICSC)*, 83-86.

Gardner, J. and Xiong, L. (2009). Automatic Link Detection: A Sequence Labeling Approach. ". *Proceeding of the 18th ACM conference on Information and knowledge management*, 1701-1704.
Geva, S. et al. (2009). Link Discovery in the Wikipedia. *Pre-Proceedings of INEX 2009*, Australia,326-333.
Granitzer, M. (2009). Context based wikipedia linking. *Advances in Focused Retrieval*, 354-365.
Green, S.J. (1998). Automated link generation: can we do better than term repetition? *International Conference on Information Technology: Coding and Computing (ITCC 04)*, 30(1-7), 75-84.
Isa, D. et al. (2009). Using the self-organizing map for clustering of text documents. *Expert Systems With Applicaions*. 36(5), 9584-9591.
Itakura, K.Y. et al. (2011). Topical and Structural Linkage in Wikipedia. *Proceedings of the 33rd European conference on Advances in information retrieval*. 460-465.
Kaski, S. et al. (1998). Websom-self-organizing maps of document collections. *Neurocomputing*, 21, 101–117.
Knoth, P. et al. (2010). Automatic generation of inter-passage links based on semantic similarity. *In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 590-598.
Kohonen, T. (1997). *Self-organizing maps*. Springer.
Landow G.P. (2006). *Hypertext 3.0: critical theory and new media in an era of globalization*. Johns Hopkins University Press.
Lee, C.H. and Yang, H.C. (1999). A web text mining approach based on self-organizing map. *In: Proceedings of the ACM CIKM'99 second workshop on web information and data management*, 59–62.
Liu, Y. et al. (2008). ConSOM: A conceptipnal self-organizing map model for text clustering. *Neurocomputing*. 71(4-6), 857-862.
Matharage, S. et al. Fast Growing Self Organizing Map for Text Clustering. *Neural Information Processing Lecture Notes in Computer Science*, 406-415.
Milne, D and Lan, H.W. (2008). Learning to link with Wikipedia. In Proceeding of the 17th ACM conference on Information and knowledge management. California, USA, 509-518.
Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. *First International Conference on MAchine Learning*
Rauber, A. and Merkl D. (1999). Using self-organizing maps to organize document archives and to characterize subject matter: How to make a map tell the news of the world. *In: Proceedings of the 10th international conference on database and expert systems applications*, 302–311.
Salton, G. and Buckley, C. (1989). *On The Automatic Generation of Content Links In Hypertext*. Department Of Computer Science, Cornell University, Ithaca, NY 14853-7501.
Singhal, A. and Salton, G. (1995). Automatic Text Browsing Using Vector Space Model. *Proceedings of the Dual-Use Technologies & Applications Conference*. 318-324.
Yang, H.C. and Lee, C.H. (2005). A text mining approach for automatic construction of hypertexts. *Expert Systems with Applications*. 29(4), 723-734.
Yang, H.C. and Lee, C.H. (2012). A Novel Self-Organizing Map for Text Document Organization. *Innovations in Bio-Inspired Computing and Applications*, 39-44.
Zeng, J. and Bloniarz, P. (2004). From Keywords to Links: an Automatic Approach. *International Conference on Information Technology: Coding and Computing (ITCC 04)*, 283-286.

Zhang, J. and Kamps, J. (2009). Link detection in XML documents: What about repeated links? *In: Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*.59-66.