ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Analysis of Implicit Relationship Using Association Link Network on Wikipedia

A. Abu Ayyoub Ansari[1], Dr. S. Thiruvenkadam[2]

Student, M.E, Dept. of CSE, P. A. College of Engineering & Technology, Pollachi, TamilNadu, India [1].

Professor and Head, Dept. of IT, P. A. College of Engineering & Technology, Pollachi, TamilNadu, India[2].

ABSTRACT: Searching web pages containing keywords that have grown in this period, while recent research is based on knowledge information. The proposed system focus on measuring relationships between pairs of keywords in Wikipedia, those keywords can be called as objects. Generalized Maximum flow method is used to reveal the objects that having high degree. It finds the similarity between the pairs of objects in Wikipedia using Association Link Network method. Association Link Network is a kind of semant0ic link network designed for establishing the associated relations among various web resources. ALN is mainly for finding the indirect relationship between the pair of object at the resource level. One of the methods used for finding similarity is Element Fuzzy Cognitive Map, is used to measure the association link at the resource level.

KEYWORDS: Association Link Network, Elements Fuzzy Cognitive Map, Generalized Maximum flow Method.

I. INTRODUCTION

Association Link Network (ALN) is the discovery of the associated relations among resources. Based on the associated relations at keyword-level, algorithm is discovered for finding the associated relations at resource-level. ALN is a semantic link network, for designing the semantic relationship at the resource level [9].

To represent various web resources some of the methods are followed, as Resource Description Framework (RDF), Ontology Inference Layer (OIL), frame-based Simple HTML Ontology Extensions (SHOE), and Web Ontology Language (OWL), etc. Automatically building of semantic link at the resource level is tedious so Element- Fuzzy Cognitive Mapping (E-FCM) is developed [9].

Element- Fuzzy Cognitive Mapping (E-FCM) shows how the causal concept affects the another concept to some other degree. Causal concept is a virtual word includes events, values, moods, trends, goals [12], etc. The wide recognition of FCMs as a promising modelling and simulation methodology for complex systems, characterized by abstraction, flexibility and fuzzy reasoning, promoted the research on new concepts in this area.

Associated relation exists between not only keywords but also E-FCMs, which means that a resource may be a cause or effect of another resource. For example, if a user browses “obese people face bigger cancer risk”, then he/she may possibly want to browse some resources about “weight loss [9]”.

In existing system generalized maximum flow reflects the three factors such as distance, connectivity, co-citation and it does not underestimate objects having high degree. It measures the strength of a relationship by using the elucidatory object. In existing system keyword level relationship only established, indirect relationship between the keywords does not exist.

The proposed system is developed, that finds the relationship between the pairs of objects (keywords) in Wikipedia using Association Link Network method. And suggesting the links for the new keywords based on the semantic relationship between the keywords.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

II. RELATED WORKS

Web mining is used to discover the pattern from the Web. The information collection should be done from the recorded source. From this collected information more knowledge can be gained for particular issue and analysis the issue. According to analysis target, web can be divided into three different types, Web usage mining, Web content mining and Web structure mining.

Web usage mining is the process of extracting useful information from server logs. One of the applications of data mining techniques is web usage mining for discovering interesting pattern from the web.

The similarity between words and phrases based on information distance and Kolmogorov complexity is measured in [10]. The world-wide-web as database and Google as search engine is used. The method is also applicable to other search engines and databases. A method is constructed to automatically extract similarity, the Google similarity distance, of words and phrases from the world-wide-web using Google page counts. WordNet database is used as an objective baseline against that to judge the performance of our method. A massive randomized trial in binary classification is used to support vector machines to learn categories based on our Google distance.

The semantic relationship is measured in [2]. Measure of semantic relatedness can be done by using the hyperlink structure on behalf of the categorical hierarchy or textual content. Semantic relationship for various terms need some knowledge even for general relationship, some common knowledge must be required. A novel approach is proposed to computing semantic relatedness with the aid of Wikipedia. Wikipedia Link-based Measure forms a compromise between these very different methods by utilizing Wikipedia's network of inter-article links, rather than the comparatively small category hierarchy used by the former system, or the full textual content used by the latter.

The use of a semantic relatedness measure between words is explored in [4], that uses the Web as knowledge source. This measure exploits the information about frequencies of use provided by previous search engines. A new semantic relatedness measure is defined among ontology terms. The relatedness among word senses expressed as terms from any ontology, instead of plain words. Wikipedia-based methods include an evaluation with some of the benchmarks.

RiTa is a Toolkit for Computation Literature and introduced in [3]. It is a suitable for open-source components, tutorials, then provide the support for a range of tasks related to the practice of creative writing in programmable media. RiTa also covers a range of computational tasks related to literary practice, including text analysis, generation, display and animation, text-to-speech, text-mining, and access to external resources such as WordNet.

RiTa provides an extensible end-to-end programming framework for assignments, projects, and class demonstrations in computational literature classes. It is well-documented, easy to learn, and simple to use. There are several frameworks for teaching natural language and computational linguistics, Natural Language Processing of these focuses on the specific requirements of an arts context

.

Association thesaurus that covers almost 1.3 million concepts in [5] and its accuracy is proved in detailed experiments. A scalable method for constructing an association thesaurus from Wikipedia based on link co-occurrences is developed. Link co-occurrence analysis is more scalable than link structure analysis because it is a one-pass process. The integration method of tfidf and link co-occurrence analysis is proposed. The integration of tfidf achieved higher accuracy than using only link co-occurrences.

A network concept from Wikipedia documents using a random walk approach is used to compute distances between documents in [9]. Three algorithms for distance computation are considered. The methods are hitting commute time, personalized page rank, and truncated visiting probability. In parallel, four types of weighted links in the document network are considered. The actual hyperlinks, lexical similarity, common category membership, and common template use. The resulting network is used to solve three benchmark semantic tasks - word similarity,

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

paraphrase detection between sentences, and document similarity by mapping pairs of data to the network, and then computing a distance between these representations.

III. CATEGORY GROUPING

Web page consists of many words like phrase, noun, verb, etc and some important words are taken as keywords. Among those important keywords, redundant keywords are removed and used for finding the relationship between the keywords [3].

Wikipedia pages are collecting from the wiki dump, which are available in web. Collect more number of Wikipedia pages. Keywords are extracted from the Wikipedia pages. For the keyword extraction pre-processing is taken place [5]. From the extracted keyword relationship is determined by using RiTa WordNet, cluster the keywords which are having the threshold of about 0.5.

Category grouping is the process of grouping similar keywords from the pre-processed keywords. Comparison between the keywords is taken place before clustering. RiTa Word Net is used for the comparison, it is nothing but application toolkit for finding the relationship between the two lemma, that is start and the end words (query and the keywords) [3]. RiTa WordNet can work under the WordNet library, it ingress word net library as the external direct source. Working principle of the RiTa is if a word is given for finding similarity first it will form the tree structure in the WordNet and take the query as the root of the tree structure.

Then consider the tree structure as the key point for finding relationship with respect to distance. The relationship is based on distance because we are mainly concentrating on the explicit and implicit relationship between two keywords. Finally relationship values of each and every keyword are determined. Threshold should be fixed for clustering the related object. The keywords which are having threshold values of about 0.5 are grouped.

Finding Relationship Between The Keywords

Step 1: A category grouping representing a concept might have descendant categories each representing its sub concept.
Step 2: For grouping similar concept RiWordNet is used.
Step 3: Provides library support for application and applet access to WordNet.

Finds the relationship between the two keywords locate node cp, the common parent of the two lemmas,

1. calculate minDistToCommonParent
2. calculate distFromCommonParentToRoot
3. (minDistToCommonParent/(distFromCommonParentToRoot+minDistToCommonParent))

IV. GAIN FUNCTION

After category grouping, grouped keywords are taken and label the nodes. If a node (keyword) which is directly connected to the source (s) to destination (v) is explicitly related, otherwise they are implicitly related.

Consider the edge is (u,v) which depends on a distance function d(e): if $u \in S \wedge v \in T$ or $u \in T \wedge v \in S$, then $d(e) = 0$; if $u \in S \wedge v \in S$ or $u \in T \wedge v \in T$, then $d(e) = 1$ [11]. Basically d(e) can be calculated by connecting edge of the node from root. That is if a node directly connected with root then distance function should be $d(e) = 0$, if a node connected to root through another then distance function $d(e) = 1$, else two nodes are connecting to reach root means distance function $d(e) = 2$, then continues till it reaches to leaf node[11].

By using distance function, gain function for the each and every node can be calculated, we introduced two parameters α and β,

$$\gamma(e) = \alpha * \beta^{d(e)}, \quad 0 < \alpha < 1, \; 0 < \beta \leq 1 \qquad (1)$$

For calculating reverse gain function we introduce another parameter λ as,

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

$$rev(e) = \lambda \times \gamma(e),\ 0 \leq \lambda \leq 1 \qquad (2)$$

Where, λ is used to adjust the importance of the reversed edge.

V GENERALIZED MAXIMUM FLOW ALGORITHM

Generalized maximum flow problem is similar to the classical maximum flow problem except that each and every flow of edge has a gain function γ (e) > 0. We introduce the term capacity of the edge μ (e), over all weight of the each and every edge. Consider the condition that flow of edge f (e) ≤ μ (e) [11].

Let generalized network consists of graph G = (V, E, s, t, μ, γ), information network (V, E) with source s ϵ V, the destination t ϵ V [8]. Flow value can be defined as total amount of f arrived at destination t. Qualitative ascertain the claim that our method can reflect the three representation concepts; they are distance, connection, co-citation.

Distance is length of the shortest path, this shortest path represent the strongest relationship between the keywords. The hitting time from vertex s to vertex t [10], that is from source to destination is defined as expected number of random walk starting from s to v at the first visit. The Truncating Hitting time (THT) to compute a largest path connecting two vertices s to v that is source and destination [8].

Connectivity is more precise the vertex connection between the source s to destination v on the network is strong if one of the node or edge is disconnected the connection will not occurs. If the connection between s to v is large then consider that relationship between source and destination is very strong. If the connection between s to v is large, consider the capacity value 1 that is capacity =1. The capacity calculation according to the flow value between the connections [1]. Value of flow is mainly depends on the capacity of the vertex between s to v. Consider the factor maximum flow analysis, it is mainly depend on the flow value of the network. If the distance is of maximum flow is decreases.

Co-citation can be done by measuring number of object connecting to and fro to the both source and destination. The strength can be determined by using number of object connected to both nodes. Consider an example for strength of the relationship can be considered by using the counting of number of Wikipedia pages. For finding strength of the relationship, gain function places an important role. In that reverse path cannot be taken place for finding to and fro connection so it is not used in 3-hop network [2]. So sometimes it cannot be used here. But consider the flow of network that is edges can give the reverse path value as 0 and compute the network in the efficient form.

VI FUZZY COGNITVE MAPPING

Fuzzy Cognitive Map (FCM) is a symbolic representation for the describing the complex system. Fuzzy Cognitive Maps are fuzzy signed graphs with feedback. Its nodes-concepts C_i and interconnections e_{ij} between concept C_i and concept C_j [12]. Weight be w_{ij}.

The computation of new value the pervious value concept can be represented by the coefficient k_2 and the k_1 expresses the influence of the interconnected concepts in the configuration of the new value of the concept A_i. This new formulation assumes that a concept links to itself with a weight $w_{ii} = k_2$ [7].

Each node-concept represents one of the key-factors of the system. Relationships between concepts have three possible types; either expresses positive causality between two concepts ($Wij > 0$) or negative causality ($Wij < 0$) or no relationship ($Wij = 0$). The value of Wij indicates how strongly concept Ci influences concept Cj. The sign of Wij indicates whether the relationship between concepts Ci and Cj is direct or inverse [12]. The direction of causality indicates whether concept Ci causes concept C_j, or vice versa. These parameters have to be considered when a value is assigned to weight Wij. A new formulation for calculating the values of concepts of Fuzzy Cognitive Map is proposed [12].

$$A_i^t = f(\sum_{j=1}^{n} A_j^{t-1} W_{ij} + A_i^{t-1}) \qquad (3)$$

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

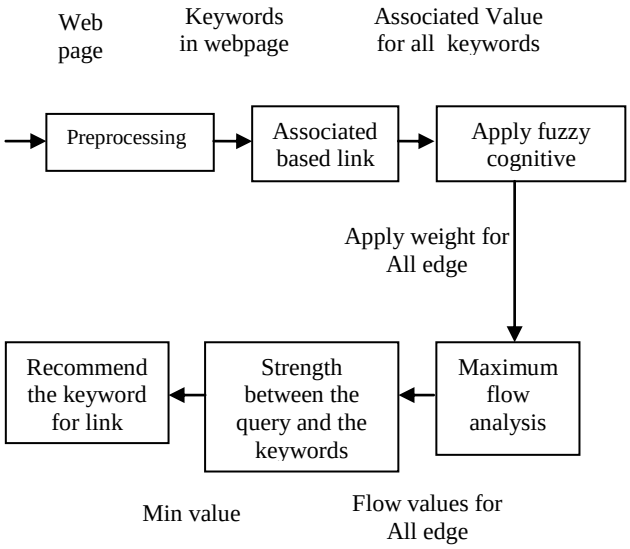

Figure 1.System Architecture

Element Concept For Fuzzy Cognitive Mapping

Step 1: Definition of the initial vector A that corresponds to the elements-concepts identified by experts' suggestions and available knowledge.
Step 2: Multiply the initial vector A with the matrix *W* defined by experts.
Step 3: The resultant vector A at time step *k* is updated using function threshold ' *f* '.
Step 4: This new vector is considered as an initial vector in the next iteration.
Step 5: Steps 2–4 are repeated.

VII. ASSOCIATION LINK NETWORK (ALN)

ALN is the discovery of the associated relations among resources. Based on the associated relations at keyword-level, we develop a discovery algorithm of associated relations at resource-level [6].

Associated relation exists between not only keywords but also E-FCMs, which means that a resource may be a cause or effect of another resource. For example, if a user browses "obese people face bigger cancer risk," then he/she may possibly want to browse some resources about "weight loss." It is obvious that the former resource is a cause of the latter resource.

The prior knowledge of topics composed by association rules at keyword-level is stored in a matrix. In addition, each element of the prior knowledge matrix indicates the weight of the associated relation at keyword-level in a domain. The association weight from resource A to resource B denoted AwT(A,B) as is the sum of the associated relations whose causal element concepts are in resource A and effect element concepts are in resource B [4].

For example, if we want to get the association weight from E-FCM1 to E-FCM2, we obtain CC4 = [0, 0, 1, 1, 0, 0, 0] first. Then, we can get the index of the element concepts which equal to 1 in CC (4). Second, sum up all the values of each column in the prior knowledge matrix of resource 1 represented by and get the vector [0.7, 0.1, 0, 0.1, 0, 0.1, 0.7], as shown in Table III. Third, perform "AND" operation on vector [0.7, 0.1, 0, 0, 0.1, 0.2, 0.3] and CC (3). The result is [0.7, 0.1, 0, 0, 0.1, 0.2, 0.3].

Finally, sum up all elements in the vector together and get [9]

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

AwT = 0.7+0.1+0 =0.8 CC (4) = 0.8

The associated relation between two resources is determined by the pairs of associated relations with keyword-level.

Indirect Relationship on Association Link Network
Step 1: Collect the FCM value and Relationship value for all keywords.
Step 2: Generate effective weight for those keywords.
Step 3: Compare effective weight and original weight for those keywords, if the original weight and the effective weight are same.
Step 4: Perform AND operation and add the values and make this value as associated value for those keywords

VIII. RESULT

This system is implemented by Associated Link Network (ALN). ALN is a kind of semantic link network, which is designed to establish associated relations among various resources aiming at extending the loosely connected network of no semantics to an association-rich network. ALN is mainly for determining the indirect relationship between the pairs of the object (keyword). The aim of this system is to increase the number of the keywords, by finding the strength of the relationship between the keywords.

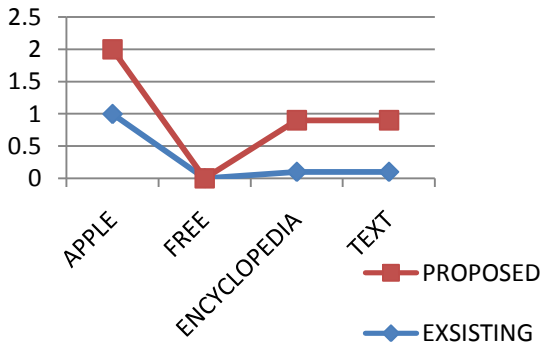

Figure.2 Strenght graph

The above figure 2 shows the strength of the keyword, the strength of the word varies because of using Association Link Network, which indicates the indirect relationship between the keywords.

IX. CONCLUSION

This system is developed for measuring the strength of a relationship between two objects on Wikipedia. And also aim to extending the link of the keywords in Wikipedia. By using the Associated Link Network (ALN) method associated resources has been developed based on the association rules between keywords. It extends the associated relation from keyword level to resource-level.

Compare the Generalized Maximum flow analysis and Associated Link Network, from the number of keywords suggested by more than 25% of the users. This shows that the number of keywords suggested in existing (20%) is increased to **92%** of proposed system for recommending link in Wikipedia page.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Some future challenges remain. The strength of the keyword may get increased by using some other algorithm which is used in the neural networks. And also for large keywords efficiency should be increased.

REFERENCES

[1] Klusch, Matthias, and Patrick Kapahnke (2012), "Web Semantics: Science, Services and Agents on the World Wide Web", Web Semantics: Science, Services and Agents on the World Wide Web 15, pp.1-14.
[2] Witten, I., and David Milne (2008), "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links ", Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA.
[3] Howe, Daniel C (2009), "RiTa: creativity support for computational literature."Proceedings of the seventh ACM conference on Creativity and cognition".
[4] Gracia, Jorge, and Eduardo Mena (2008), "Web-based measure of semantic relatedness ", Web Information Systems Engineering-WISE 2008. Springer Berlin Heidelberg, pp.136-150.
[5] Ito, Masahiro, et al. (2008), "Association thesaurus construction methods based on link co occurrence analysis for wikipedia", Proceedings of the 17th ACM conference on Information and knowledge management. ACM.
[6] Mashinchi, M. Reza, and Ali Selamat (2008), "Constructing a customer's satisfactory evaluator system using GA-Based fuzzy artificial neural networks", World Appl. Sci. J Vol.5.4, pp.432-440.
[7] Perusich, Karl, and Michael D. McNeese (2006), "Using fuzzy cognitive maps for knowledge management in a conflict environment ", Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, Vol.36.6, pp.810-821.
[8] Zhang, Xinpeng, Yasuhito Asano, and Masatoshi Yoshikawa (2013), "A generalized flow basedmethod for analysis of implicit relationships on wikipedia", pp. 1-1.
[9] Luo, Xiangfeng, et al. (2011), "Building association link network for semantic link on web resources",Automation Science and Engineering, IEEE Transactions Vol.8.3, pp.482-494.
[10] Cilibrasi, Rudi L., and Paul MB Vitanyi (2007), "The google similarity distance", Knowledge and Data Engineering, IEEE Transactions Vol.19.3, pp.370-383.
[11] Bollegala, Danushka T., Yutaka Matsuo, and Mitsuru Ishizuka (2009), "Measuring the similarity between implicit semantic relations from the web" ,Proceedings of the 18th international conference on World wide web. ACM.
[12] Stylios, Chrysostomos D., and Peter P. Groumpos. "Fuzzy cognitive maps in modeling supervisory control systems ." Journal of Intelligent and Fuzzy Systems.