Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 9 (4): 794-798, 2009
ISSN 1812-5654

Knowledge Acquisition from Textual Documents for the Construction of Medicinal Herbs Domain Ontology

[1]I. Zaharudin, [1]S.A. Noah and [2]M.M. Noor
[1]Department of Information Science, Faculty of Information Science and Technology,
[2]School of Bioscience and Biotechnology, Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, 43650 Selangor, Malaysia

Abstract: In this study a semi automatic acquisition of domain relevant terms from digital documents in e-newspaper related to Malaysian medicinal herbs is presented. This study proposes (1) TFIDF-based term classification method for acquiring single word terms, (2) recognition of multi-word using TerMine software to acquire multiword terms and (3) Hearst's methodology of acquiring semantic relationships of hyponym. The results show the benefits of using these methods in selecting relevant terms from domain specific corpus. From this study it is believed that the combination of these three methods might be helpful to select relevant terms as well as minimize the effort to discard irrelevant terms manually from wide collection of terms from the corpus.

Key words: Knowledge management and extraction, Malaysian medicinal herb, semantic web, natural language processing, biomedical, knowledge engineering

INTRODUCTION

Knowledge has been distinguished from data and information in two different ways (Becerra-Fernandez *et al*., 2004). The first view considered knowledge to be as the highest level in the hierarchy compared to information and data. According to this view, knowledge refers to information that enables action and decisions, or information with direction. Hence knowledge is intrinsically similar to information and data, although it is the richest and deepest of the three and consequently also the most valuable. Based on this view data refers to bare facts void of context. Whereas information is considered as data in context and knowledge is information that facilitates action. Although the aforementioned simplistic view about knowledge may not be completely accurate since it does not fully explain the characteristic of knowledge, the second view defined knowledge as an area as justified beliefs about relationships among concepts relevant to that particular area. This take into consideration that knowledge is intrinsically difference from information and also richer and more detail set of facts. This definition is consistent with Nonaka and Takeuchi (1995), which define knowledge as a justified true belief. Whereas Wigg (1999) views knowledge as fundamentally different from data and information since knowledge consist of truths and beliefs, perspectives and concepts, judgments and expectation, methodologies and know-how and is possessed by human, agents, or others active entities and is used to receive information and to recognize and identify; analyze, interpret and evaluate; synthesize and decide; plan, implement, monitor and adapt to act more or less intelligently. In other words, knowledge is used to determine what a specific situation means and how to handle it.

Ontology is a kind of knowledge which was historically introduced by Aristotle. Ontology recently has become a topic of interest in computer science. Ontology provide a shared understanding of a domain of interest to support communication among human and computer agents, typically being represented in a machine processable representation language (Haase and Sure, 2004). According to Staab *et al*. (2001) ontology is an explicit formal specification of terms, which represent the intended meaning of concepts, in the domain and relations among them. An ontology can also be defined as a formal structuring of knowledge. In its purest form, it is meant to represent reality, which means some part of the world as we currently understand or interpret it. An ontology consists of universals (also referred to variously as entities, classes, concepts, types and terms) and the relationships between them (Smith, 2003; Smith *et al*., 2003). A universal is simply a type, or

Corresponding Author: Shahrul Azman Noah, Department of Information Science, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43650 Selangor, Malaysia
Tel: +60389216178 Fax: +60389256732

category, of things in the real world. Universals are often divided into two main subtypes: continuants (things that exist) and occurrences (things that occur in time, or events). There are many ways ontology can be represented, spanning from a simple concept into a complex axiom. However, a concept hierarchy or taxonomy has been the favorite representation by many developer due its simplicity but yet representable.

Ontology is considered as the backbone of many current applications such as knowledge-based systems, knowledge management systems and semantic web applications. One of the important tasks in the development of such systems is knowledge acquisition. Conventional approaches to knowledge acquisition are mainly from interviewing domain experts and subsequently modeled and transform the acquired knowledge into some forms of knowledge representation technique. However, huge amount of knowledge is currently embedded in various academic literatures and has the potential to be exploited for knowledge construction. The main inherent issue is that such knowledge is highly unstructured and difficult to be transformed into meaningful model. Although a number of automated approach to acquiring such knowledge has been proposed by Alani *et al.* (2003) and Cimiano *et al.* (2005) their success are yet to be seen. Such approaches also have only been tested on general domain and scientific domains such as the medicinal herbs domain are yet to be explored. While automated approach seems to offer promising solutions, human still play an important role in validating the correctness of the acquired knowledge particularly in scientific domain. This study, therefore, proposed a semi automated approach for acquiring domain knowledge of the medicinal herbs domain from academic literature. A combination of Natural Language Processing (NLP) and statistical techniques have been employed for extracting concept terms from the literature. The Hearst heuristic rules on the other hand have been adopted for building the domain taxonomy.

MATERIALS AND METHODS

According to Fuller *et al.* (2004) there is an overwhelming increase in the biomedical literature. The advances in the biomedical sciences need the development of ontology to assist user understand the developments in one's own area of specialization and also enable them to learn quickly about developments in related and unrelated subject areas. The ontology may also helpful to fulfill the need to uncover information present in large and unstructured bodies of text, commonly referred to as non-interactive literatures (Swanson and Smalheiser, 1997); i.e., literatures that do not cite each other but which, nevertheless, together present useful new information.

The main objective of this study is to explore the possibility of extracting semantic information or knowledge from medicinal herbs literature and to represent it in the form of some ontology structure. As ontology can be represented in many forms starting with a simple list of concepts till the complexity of logic structure, we focus on representing it in terms semantic net i.e., concepts with semantic links. The study focused on our initial findings based upon current approaches of extracting knowledge from unstructured documents as previously described. The study was conducted between the Knowledge Technology Research Group and the Reproductive and Developmental Biology Research Group of the Universiti Kebangsaan Malaysia. It is currently conducted at the premises of the aforementioned research groups. The study is currently in progress and expected to be completed in the end 2009.

Research approach: This study is at the preliminary stage in constructing the domain specific medicinal herbs ontology. This study involved the gathering of medicinal herbs documents and the analysis of the documents using NLP tools and statistical analysis. This study applied a semi-automatic approach in collecting relevant terms in medicinal herb documents to be inserted as ontological term. For the initial or pilot stage, 18 documents related to medicinal herbs have been collected from Malaysia digital newspaper archive. The collection has been analyzed to extract relevant terms that can be used as ontological terms. The analyzing processes involved are simplified into the following stages (Fig. 1).

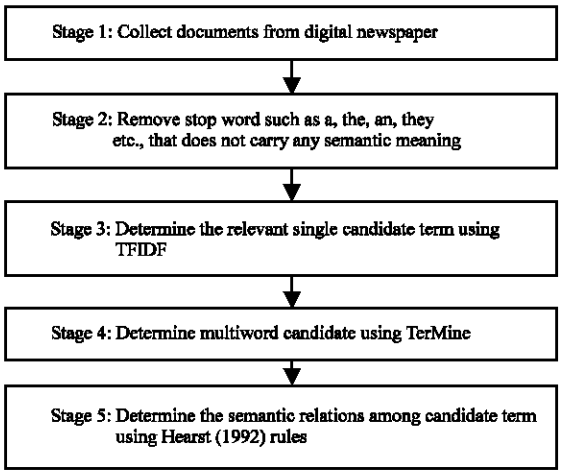

Fig. 1: The stages of the document analyzing processes

RESULTS AND DISCUSSION

As this initial study only 18 documents were used and all the terms extracted were considered for the analysis, whatever their frequency and c-values are. Though realistically only terms with higher weight should be considered once we have huge collections with massive terms. For the semantic relation, only hyponym (i.e., IS-A), which is based on the study of Hearst (1992) has been extracted.

Single terms selection: A simple program to eradicate stopword was developed using Python programming and acquired domain relevant single-word terms by deploying TFIDF-based term classification method based on the following equation:

$$w_{ij} = tf_{ij} \times \log\left(\frac{D}{df_i}\right)$$

where, tf_{ij} refers to the No. of term, i in document j, N as the No. of document in a collection and dfi as the No. of documents in a collection containing term i.

In this study 1484 different terms were found. The term that has highest TFIDF value was ranked as the most relevant term in this domain while the lower TFIDF value has the lower relevancy of the term in this domain. Table 1 shows the first 20 single terms ranked according to the TFIDF weight.

Multiword selection: For multiwords or phrases selection the TerMine software provided online by The National Center for text Mining (NaCTem) was used and listed out 20 highest C-values obtained from the corpus analysis (Table 2). According to Frantzi *et al.* (2000) C-value enhances the usual statistical measure of frequency of term occurrence, making it sensitive to the nested terms. The C-value is obtained from the following equation:

$$C-value(a) = \log_2 |a| \cdot ((f(a) - \frac{1}{P(Ta)} \sum_{b \in Ta} f(b))$$

where, a is the candidate term, f(.) is its frequency of occurrence in the set of documents (corpus), Ta is the set of candidates that contain a, P (Ta) is the number of these candidate terms.

These multiword semantic relations will later be refined in the next process which will be presented later.

Semantic relationship identification: The Hearst (1992) technique was applied to verify the hyponym semantic relation and initially found 29 hyponym (IS-A) relations from this collection (Table 3). The Hearst rules for detecting hyponym from text include:

(1a) NPo ..... such as {NP1, NP2 . . . . (and /or)} NP,,
are such that they imply
(1b) for all NP,, 1 < i< n, hyponym(NPi, NPo)
Thus from sentence it can be concluded that
Hyponym (Barn bare n dang, bow lute)

(2) such NP as {NP ,}* {(or [and)} NP
... works by such authors as Herrick,
Goldsmith and Shakespeare.
Hyponym (author, Herrick),
Hyponym (author, Goldsmith),
Hyponym (author, Shakespeare)

Table 1: List of terms based on TFIDF

Single term	Frequency	TFIDF
Biotechnology	59	17.761
Tongkat	21	16.341
Ali	18	14.007
Company	17	11.105
Herbal	23	10.974
Traditional	25	10.254
Hovid	8	10.042
Cinnamon	7	8.787
Cells	9	8.588
Products	31	7.913
Medicine	14	7.788
Ministry	14	7.788
Alternative	10	7.782
Industry	30	7.658
Ins	8	7.634
Natural	16	7.634
Felda	6	7.532
India	6	7.532
Market	25	7.526
Quality	11	7.185

Table 2: List of multiword based on C-values

Multiword	C-values
Tongkat Ali	19.81250
Prime minister Datuk Seri Abdullah Ahmad Badawi	8.42206
Herbal product	8.00000
Natural product	7.00000
Alternative medicine	6.00000
Biotechnology industry	6.00000
Traditional medicine	6.00000
Herbal product industry	5.33985
Tongkat Ali coffee	5.33985
Prime Minister	5.14286
Kuala Lumpur	5.00000
Traditional herb	5.00000
National biotechnology policy	4.75489
Malaysian biotechnology corporation sdn bhd	4.64386
Akademi technology park Malaysia	4.00000
Business times	4.00000
Mohd Nazlee	4.00000
Private sector	4.00000
United state	4.00000
World bank	4.00000

Table 3: List of semantic relationship on hyponym

No.	Types of semantic relations
1	HYPONYM (herbs, cloning of plant)
2	HYPONYM (Hovid, local pharmaceutical companies)
3	HYPONYM (processing dried plant materials to herbal powders, SMEs low value-added activities)
4	HYPONYM (fermentation to produce medicinal tonics, SMEs low value-added activities)
5	HYPONYM (grants, biotechnology-related areas incentives)
6	HYPONYM (loans, biotechnology-related areas incentives)
7	HYPONYM (pioneer status, biotechnology-related areas incentives)
8	HYPONYM (investment tax incentives, biotechnology-related areas incentives)
9	HYPONYM (herbal natural industry, biotech sub-sectors)
10	HYPONYM (Gold Medal at the 34th Geneva International Exhibition of Inventions, accolades)
11	HYPONYM (New Techniques and Products 2006 for the development of Deng Kang Lu, an anti-mutation product made from traditional herbs, accolades)
12	HYPONYM (Silver Medal for its diabetes product Insupro Forte which contains plant insulin, accolades)
13	HYPONYM (China, countries)
14	HYPONYM (US, countries)
15	HYPONYM (Japan, countries)
16	HYPONYM (screening, new services and technologies)
17	HYPONYM (early diagnosis of specific diseases, treatment and interventions, new services and technologies)
18	HYPONYM (developing a network or nexus , industry)
19	HYPONYM (centres of excellence from existing institutions around the country, known as Bionexus Malaysia, industry)
20	HYPONYM (improving responsiveness to patient needs and expectations, quality of the soft component of care)
21	HYPONYM (ensuring staff competency and capability, quality of the soft component of care)
22	HYPONYM (government-linked companies, institutions)
23	HYPONYM (10 year tax exempt status, incentives)
24	HYPONYM (Type II diabetes, disease)
25	HYPONYM (palm oil, commodities)
26	HYPONYM (rubber, commodities)
27	HYPONYM (cocoa, commodities)
28	HYPONYM (herbs, commodities)
29	HYPONYM (high risks surrounding the industry, cost for research and development)

(3) NP {, NP} * {,} or other NP
Bruises, wounds, broken bones or other injuries . . .
Hyponym (bruise. injury),
Hyponym (wounds, injury),
Hyponym (broken bone, injury)

(4) NP {, NP}* {,} and other NP
... temples, treasuries and other important civic buildings.
Hyponym (temples, civic building)
Hyponym (treasury , civic building)

(5) NP {,} including {NP * {or / and} NP
All common-law countries, including Canada and England ...
Hyponym (Canada, common-law country),
Hyponym (England, common-law country)

(6) NP {,} especially {NP ,}* {or] and} NP
. . . most: European countries, especially France, England and Spain.
Hyponym (France, European country),
Hyponym (England, European country),
Hyponym (Spain, European country)

Table 3 shows 29 hyponym relationships detected in the collection using Hearst rules.

CONCLUSION

In this study a semi-automatic approach for extracting domain relevant terms and relations among them was presented. The TFIDF based term extraction has shown that it has the potential for the extraction of single word terms; whereas the c-value method (as implemented in the TerMine software) prove to be beneficial in extracting multiword terms. On the other hand employing the Hearst technique has shown promising results in extracting the hyponym relationships among extracted concepts. From this study, it is believed that the combination of these three methods might bring more option to select relevant terms and minimize the effort to discard irrelevant terms manually from wide collection of terms from the corpus. This will help to minimize uses of domain expert which is sometimes become a constraint in the development of a specific domain ontology. Our near future works include increasing the document collection and to employ an advance NLP technique in extracting richer semantic relationships.

REFERENCES

Alani, H., S. Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis and N.R. Shadbolt, 2003. Automatic ontology-based knowledge extraction from web documents. IEEE Intelligent Syst., 18: 14-21.

Becerra-Fernandez, I., A. Gonzalez and R. Sabherwal, 2004. Knowledge Management Challenges, solutions and Technology. 1st Edn., Prentice Hall, New Jersey, ISBN: 0131099319.

Cimiano, P., A. Hotho and S. Staab, 2005. Learning concept hierarchies from text corpora using formal concept analysis. J. Artificial Intelligence Res., 24: 305-339.

Frantzi, K., S. Ananiadou and H. Mima, 2000. Automatic recognition of multi-word terms: The C-value/NC-value method. Int. J. Digital Libraries., 3: 115-130.

Fuller, S., D. Revere, P. Bugni and G.M. Martin, 2004. A knowledgebase system to enhance scientific discovery: Telemakus. Biomedical Digital Libraries, 1: 2-2.

Haase, P. and Y. Sure, 2004. State-of-the-Art on Ontology Evolution. Institute AIFB, University of Karlsruhe. http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/SEKT-D3.1.1.1.b.pdf.

Hearst, M., 1992. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th International Conference on Computational Linguistics, 1992, Association for Computational Linguistics Morristown, NJ, USA, pp: 539-545.

Nonaka, I. and H. Takeuchi, 1995. The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation. 1st Edn., Oxford University Press, New York, ISBN: 0195092694.

Smith, B., 2003. Ontology. In: Blackwell Guide to the Philosophy of Computing and Information, Floridi, L. (Ed.). Oxford Blackwell, Oxford, ISBN: 9780631229193, pp: 155-166.

Smith, B., J. Williams and S. Schulze-Kremer, 2003. The ontology of the gene ontology. AMIA Annu. Symp. Proc., pp: 609-613. http://ontology.buffalo.edu/medo/Gene_ontology.pdf.

Staab, S., H.P. Schnurr, R. Studer and Y. Sure, 2001. Knowledge processes and ontologies. IEEE Intelligent Syst., 16: 26-34.

Swanson, D.R. and N.R. Smalheiser, 1997. An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence, 9: 183-203.

Wigg, K., 1999. Introducing Knowledge Management into the Enterprise, Knowledge Management Handbook. 1st Edn., CRC Press, Boca Raton, ISBN: 0-8493-0238-2.