The Scientific Report
of
Research Work
[01.08.2009 - 31.07.2010]

done by

Dr Rajendra Prasath R
ERCIM Fellow [2009 - 2010]

at

IDI, NTNU
Norway

under
The ERCIM "Alain Bensoussan" Fellowship Programme

Submitted to:

European Research Consortium for Informatics and Mathematics
06902 Sophia Antipolis Cedex - France

The Scientific Report of Work done

[From 01.08.2009 to 31.07.2010]

Fellow : **Dr. Rajendra Prasath. R**
ERCIM Fellow (2009 - 2010)

Host : **Dr. Pinar Öztürk**
Associate Professor, Self Organizing Systems (SOS) Group

Place : Dept of Computer and Information Science (IDI)
Norwegian University of Science and Technology (NTNU)
Sem Sælands Vei 7-9, NO - 7491, Trondheim, Norway

Period of Stay : **01.08.2009 - 31.07.2010**

I. Scientific Activity:

During my 12 months stay at NTNU, I have been actively involved in designing algorithms for extracting and analyzing the textual content towards developing the knowledge based decision support systems. I have involved in self organizing systems group and have been working in exploring the higher order term relations through distributed representations for textual content. This work to towards building efficient Textual Case Based Reasoning Systems. We have implemented random indexing in an efficient way and analyzed the effectiveness of case retrievals. Also we have applied Holographic Reduced Representations for detecting the order information of the features in the textual content. In the later part, we have applied blocking and affordance assignment to the web documents and achieved good retrieval by focusing on the affordance of the given query with the matching affordance of the cases in the case base.

Recently, as the result of our research efforts, I came up with a system namely "**TextCLOUD**" - the Textual Case Based Reasoning system - to identify higher order term relations towards discovering the knowledge in terms of past experience from the textual reports. This system could be integrated with many existing real time knowledge based decision support systems in various industries like oil well drilling, medical diagnosis and to those systems requiring knowledge discovery from textual content. This system is scalable and could process large amount of data at a faster rate to discover and represent higher order term relations. Distributed representations handle the free text content effectively than the traditional Information retrieval methods. The proposed methods are powerful in terms of its capacity in discovering higher order term associations with the given case representation. Also we have carried out certain experiments in machine learning problems like learning age and gender from stylistic variations, text categorization using external knowledge repositories. Additionally we have proposed algorithms for distributed sorting and prefix computation algorithms using message passing strategies in interconnection networks.

ROLES and RESPONSIBILITIES:

My active participation in the research activities of IDI, NTNU are listed below:

1. **Member, Self Organizing Systems (SOS) Group:**

   Responsible for research, design and development of software projects in SOS group, leading a small team of researchers. The main focus of our work revolved around the design and analysis of distributed representations for handling the extracted knowledge from textual content using various Information Retrieval, Machine Learning and Artificial Intelligence techniques.

2. **Member, Knowledge Based Systems (KBS) Group, IDI**

   Association with Knowledge Based Systems group in NTNU allows me to access the industrial data from Verdande Technology AS - A spin off company providing decision support systems to oil well drilling. This industrial partner is a suitable test bed for testing the proposed textual CBR algorithms for knowledge discovery. The extracted knowledge base, in case of anomalous event, would help better in avoiding large delays and money as well. My focus is on developing methods that detect partial cases, validate them and build the case base.

3. **Member, Discussion Forum @ SOS:**

   Discussion Forum, consisting of seven active researchers, meets every week for discussing the specific problems arising in Natural Language Processing, Information Retrieval, Machine Learning and Artificial Intelligence. In this work, SOS Group members work in association with the people from language technology group at the Dragvoll Campus of NTNU.

4. **Member, Norwegian Artificial Intelligence Society (NAIS):**

   Norwegian Artificial Intelligence Society conducts every year the national level seminar, on recent developments in AI. My main focus on this seminar is to contribute atleast a research paper from our ongoing experiments, reviewing some of the submitted papers and supporting the organization of sessions. Also we extend our support in bringing out the publication of the presented papers in the form of an edited volume.

II. Publications during the Fellowship:

BOOK:

Rajendra Prasath, Message Passing Approaches in Interconnection Networks - Towards Distributed Applications, VDM-Verlag, Germany, September 2010 (In Press), ISBN: 978-3-639-26732-7

2010:

1. **R.Rajendra Prasath** and **Sudeshna Sarkar**, Unsupervised Feature Generation using Knowledge Repositories for Effective Text Categorization, to appear in: Proc. of the 19th European Conference on Artificial Intelligence (ECAI 2010), Lisbon, 2010.
2. **Rajendra Prasath** and Pinar Öztürk, Similarity Assessment through blocking and affordance assignment in Textual CBR, Accepted in: WebCBR 2010.

3. **Rajendra Prasath**, Learning Age and Gender using Co-occurrence of Non-Dictionary Words from Stylistic Variations, Lecture Notes in Artificial Intelligence(LNAI), 6086, Springer - Heidelberg, 2010, pp. 548-553.
4. Pinar Öztürk, **Rajendra Prasath** and Hans Moen, Distributed Representations to detect Higher Order Term Correlations in Textual Content, Lecture Notes in Artificial Intelligence(LNAI), 6086, Springer - Heidelberg, 2010, pp. 744-753.
5. Pinar Öztürk and **Rajendra Prasath**, Recognition of higher-order relations among features in textual cases using random indexing, Lecture Notes in Computer Science (LNCS), Vol. 6176, Springer - Heidelberg, 2010, pp. 272-286.

2009:

6. **R.Rajendra Prasath** and **Sudeshna Sarkar**, Improving text categorization using hyperlinks in external knowledge repository, in: Proc. of the First Norwegian Artificial Intelligence Symposium(NAIS 2009) held at: NTNU, Norway, Nov. 2009, pp. 79 – 90.
7. M.Rustagi, **R.Rajendra Prasath**, Sumit Goswami and **Sudeshna Sarkar**, Learning Age and Gender of Blogger from Stylistic Variation, Lecture Notes in Computer Science 5909, Springer-Verlag, Heidelberg, 2009, pp. 205–212.

Additional Publications (in Distributed Systems):

8. R.Rajendra Prasath, An alternative time - optimal distributed sorting algorithm on a line network, in: Proc. of 6th International Conference on Networked Computing (INC2010), Gyeongju, South Korea, 2010, pp. 64-69
9. Rajendra Prasath, Algorithms for distributed sorting and prefix computation in static ad hoc mobile networks, The 2010 International Conference on Electronics and Information Engineering (ICEIE 2010) (to appear in IEEE Xplore)

Working Papers [in Progress] (Extended Versions - to be communicated to Journals)

10. Similarity Assessment with Random Indexing through Blocking and Affordance Assignment in Textual CBR (to: Data Mining and Knowledge Discovery)
11. Effects of distributed representations towards detecting higher order term correlations in textual content (to: IEEE Trans on KDE)
12. Unsupervised Feature Generation using Knowledge Repositories for effective text categorization

My research publications are described In the sequel with the abstract and citation.

DISTRIBUTED REPRESENTATIONS FOR TEXTUAL CBR:

Case Based Reasoning(CBR), an artificial intelligence technique, solves new problem by reusing solutions of previously solved similar cases [1]. In conventional CBR, cases are represented in terms of structured attribute-value pairs. Acquisition of cases, either from domain experts or through manually crafting attribute-value pairs from incident reports, constitutes the main reason why CBR systems have not been more common in industries. Manual case generation is a laborious, costlier and time consuming task. Textual CBR (TCBR) is an emerging line that aims to apply CBR techniques on cases represented as textual descriptions. Similarity of cases is based on the similarity between their constituting features. Conventional CBR benefits from employing domain specific knowledge for similarity assessment [2]. Correspondingly, TCBR needs to involve higher-order relationships between features, hence domain specific knowledge. In addition, the term order has also been contended to influence the similarity assessment. We presented the method namely random indexing [14, 15] in which features and cases are represented using a distributed representation paradigm that captures higher order relations among features as well as term order information. Also we have applied Holographic Reduced Representations for detecting the order information of the features in the textual

content [10, 12]. Experimental results shows that random indexing is a fine alternative to Latent Semantic Indexing which is neither incremental nor a compact representation [3, 4].

This work, presented in **RSCTC 2010** Conference, is cited as follows:

> Pinar Öztürk, **Rajendra Prasath** and Hans Moen, Distributed Representations to detect Higher Order Term Correlations in Textual Content, Lecture Notes in Artificial Intelligence(LNAI), 6086, Springer - Heidelberg, 2010, pp. 744-753 [The Extended version of this paper is in progress]

TWO - STAGE MODEL FOR TEXTUAL CBR:

In this work, we envisage retrieval in textual case-based reasoning (TCBR) as an instance of abductive reasoning. The two main subtasks underlying abductive reasoning are 'hypotheses generation' where plausible case hypotheses are generated, and 'hypothesis testing' where the best hypothesis is selected among these in sequel. The central idea behind the presented two-stage retrieval model for TCBR is that recall relies on lexical equality of features in the cases while recognition requires mining higher order semantic relations among features. The proposed account of recognition relies on a special representation called random indexing [9, 15], and applies a method that simultaneously performs an implicit dimension reduction and discovers higher order relations among features based on their meanings that can be learned incrementally. Hence, similarity assessment [13, 16] in recall is computationally less expensive and is applied on the whole case base while in recognition a computationally more expensive method is employed but only on the case hypotheses pool generated by recall. The main intuition is derived from Jonshon- Lindenstrauss Lemma[8]. It is shown that the two-stage model gives promising results on the experimental text collections.

This work, presented in **ICCBR 2010** Conference, is cited as follows:

> Pinar Öztürk and **Rajendra Prasath**, Recognition of higher-order relations among features in textual cases using random indexing, Lecture Notes in Computer Science (LNCS), 6176, Springer - Heidelberg, 2010, pp. 272-286. [The Extended version of this paper is in progress]

BLOCKING AND AFFORDANCE ASSIGNMENT:

It has been conceived that children learn new objects through their affordances, that is, the actions that can be taken on them. We suggest that web pages also have affordances defined in terms of the users' information need. An assumption of the proposed approach is that different parts of a text may not be equally important / relevant to a given query. Judgment on the relevance of a web document requires, therefore, a thorough look into its parts, rather than treating it as a monolithic content. We propose a method to extract valid text blocks; assign affordances to texts and then use these affordances to retrieve the corresponding web pages. The overall approach presented in the paper relies on case-based representations, similar to [5] that bridge the queries to the affordances of web documents. We tested our method on the tourism corpus and the results are promising.

This work, presented in **WebCBR 2010** Conference, is cited as follows:

> **Rajendra Prasath** and Pinar Öztürk, Similarity Assessment through blocking and affordance assignment in Textual CBR, accepted in: WebCBR 2010 [The Extended version of this paper is in progress]

UNSUPERVISED FEATURE GENERATION:

We proposed an unsupervised feature generation algorithm using the repositories of human knowledge for effective text categorization. Conventional bag of words (BOW) depends on the presence / absence of keywords to classify the documents. To understand the actual context behind these keywords, we use knowledge concepts / hyperlinks from external knowledge sources through content and structure mining on Wikipedia. Then, the features of knowledge concepts are clustered to generate knowledge cluster vectors with which the input text documents are mapped into a high dimensional feature space and the classification is performed. The simulation results show that the proposed approach identifies the associated features in the text collection and yields an improved classification accuracy.

This work, to be presented in **ECAI 2010** Conference, is cited as follows:

R.Rajendra Prasath and **Sudeshna Sarkar**, Unsupervised Feature Generation using Knowledge Repositories for Effective Text Categorization, to appear in: Proc. of the 19th European Conference on Artificial Intelligence (ECAI 2010), Lisbon, 2010. [The Extended version of this paper is in progress]

CO-OCCURRENCE BASED STYLISTIC VARIATIONS:

In this work, we attempted to report the stylistic differences in blogging for gender and age group variations using slang word co-occurrences. We have mainly focused on co-occurrence of non dictionary words across bloggers of different gender and age groups. For this analysis, we have focused on the feature, "use of slang words", to study the stylistic variations of bloggers across various age groups and gender. We have modeled the co-occurrences of slang words used by bloggers as graph based model G = (V,E) where nodes(V) are slang words and edges(E) represent the number of co-occurrences and studied the variations in predicting age groups and gender. We have used demographically tagged blog corpus from ICWSM Spinner dataset for these experiments and used Naive Bayes classifier with 10 fold cross validations. Preliminary results shows that the concurrence of slang words could be a better choice for predicting age and gender.

This work, presented in **RSCTC 2010** Conference, is cited as follows:

Rajendra Prasath, Learning Age and Gender using Co-occurrence of Non-Dictionary Words from Stylistic Variations, Lecture Notes in Artificial Intelligence(LNAI), 6086, Springer - Heidelberg, 2010, pp. 548-553.

TEXT CATEGORIZATION USING HYPERLINKS:

Traditional text categorization systems use Bag of Words (BoW) approach which unable to achieve high categorization accuracy with text documents, because categorization depends merely on the occurrence of keywords. To understand the actual context behind these keywords, it is essential to induce additional features from external knowledge sources. In this work, we have made an attempt to enhance terms in documents using hyperlink structures present in vast repositories of human knowledge, in this case, Wikipedia. At first, for each featured topic in Wikipedia, Hyperlink text vectors are extracted. Then the input text documents are analyzed and selected features from them are mapped into hyperlink text vectors that tries to generate features related to the context of text fragments in a high dimensional space. The re-represented text documents are now classified with generated hyperlink features. The simulation results show that computing associated word relations in the given text fragment, using the extracted hyperlink text vectors without explicit semantic analyzer, yields better improvements in the categorization accuracy. The classifiers used in our experiments are: Naive Bayes and k − Nearest Neighbor. The categorization accuracy is measured against the standard Reuters - 21578 dataset.

This work, presented in **NAIS 2009** Conference, is cited as follows:

R.Rajendra Prasath and **Sudeshna Sarkar**, Improving text categorization using hyperlinks in external knowledge repository, in: Proc. of the First Norwegian Artificial Intelligence Symposium(NAIS 2009) held at: NTNU, Norway, Nov. 2009, pp. 79 – 90

LEARNING AGE AND GENDER:

We reported results of stylistic differences [6] in blogging for gender and age group variation. The results are based on two mutually independent features. The first feature is the use of slang words which is a new concept proposed by us for Stylistic study of bloggers. For the second feature, we have analyzed the variation in average length of sentences across various age groups and gender. These features are augmented with previous study results reported in literature for stylistic analysis. The combined feature list enhances the accuracy by a remarkable extent in predicting age and gender. These machine learning experiments were done on two separate demographically tagged blog corpus. Gender determination is more accurate than age group detection over the data spread across all ages but the accuracy of age prediction increases if we sample data with remarkable age difference.

This work, presented in **PReMI 2009** Conference, is cited as follows:

M.Rustagi, **R.Rajendra Prasath**, Sumit Goswami and **Sudeshna Sarkar**, Learning Age and Gender of Blogger from Stylistic Variation, Lecture Notes in Computer Science 5909, Springer-Verlag, Heidelberg, 2009, pp. 205–212

III. Attended Workshops, Seminars and Conferences:

VISITING POSITIONS / EXCHANGE PROGRAMME

	Position	Institution	Period	Nature of Work
1	ERCIM Fellowship	Dept. of Computer and Information Science(IDI), Norwegian University of Science and Technology (NTNU), Norway	August 1, 2009 to July 31, 2010	Research
2	Visiting Fellow	Artificial Intelligence Research Institute (IIIA - CSIC), University of Barcelona	May 23, 2010 to June 1, 2010	Research
3	Visiting Fellow	Swedish Institute of Computer Science (SICS), Sweden	June 09, 2010 to June 16, 2010	Research

INVITED TALKS

1. Invited talk at: IIIA - **I**nstitut d'**I**nvestigació en **I**ntelligència **A**rtificial, Spanish National Research Council, Universitat Autonoma de Barcelona, Spain on May 25, 2010
2. Invited talk at: **S**wedish **I**nstitute of **C**omputer **S**cience (SICS), Kista (Stockholm), Sweden on June 14, 2010
3. DIS Seminar at: Mitt Universititet, Ostersund, Sweden during January 27-29, 2010 from the "Self Organizing Group" of IDI, NTNU

PAPER PRESENTATIONS

COUNTRIES VISITED: South Korea (INC2010), Poland (RSCTC 2010), Italy (ICCBR 2010 / WebCBR 2010), India (PReMI 2009)

University of Piemonte Orientale, ALESSANDRIA, ITALY:

1. To Present **TWO** papers at: 18th International Conference on Case - Based Reasoning (ICCBR 2010), in Alessandria, Italy during July 19 - 22, 2010

University of Warsaw, WARSAW, POLAND:

2. Presented **TWO** papers at: 7th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2010), at University of Warsaw, Poland, during June 28-30, 2010

Dongguk University, GYEONGJU, SOUTH KOREA:

3. Presented **ONE** paper at: 6th International Conference on Networked Computing (INC 2010), at Gyeongju, South Korea, during May 11 - 13, 2010

Indian Institute of Technology, NEW DELHI, INDIA:

4. Presented **ONE** paper at: 3rd Third International Conference on Pattern Recognition and Machine Intelligence (PReMI 2009) at Indian Institute of Technology, New Delhi during December 16-20, 2009

IV – Research Exchange Programme (12 months scheme)

1. Exchange visit - I **23.05.2010 - 01.06.2010**

Artificial Intelligence Research laboratory (IIIA),
Spanish National Research Council (CSIC),
Campus Universitat Autonoma de Barcelona
08193 Bellaterra, Catalonia, Spain

Core responsibilities:

(a) Exploring new feature generation mechanisms for handling Textual Case based reasoning problems
(b) Analysis of web based Textual Case Based Reasoning systems

Advisor: Prof. Ramon Lopez de Mantaras

2. Exchange visit - II **09.06.2010 - 16.06.2010**

SICS, Swedish Institute of Computer Science AB
Box 1263 (visit: Isafjordgatan 26)
S - 164 29 Kista, Sweden

Core responsibilities:

(a) Learning Age and Gender from bloggers using stylometric analysis
(b) Exploring web content for Textual Case Based Reasoning systems
(c) Discussions on distributed representation for detecting higher order term relations in textual descriptions

Advisor: Prof. Björn Gambäck