Tools to Find Connections Between Researchers – Findings from Preliminary Work with a Prototype as Part of a University Virtual Research Environment

Jim Hensman[1], Dimoklis Despotakis[2], Ajdin Brandic[1], Vania Dimitrova[2]

[1] Coventry University, UK
{j.hensman, a.brandic}@coventry.ac.uk
[2] University of Leeds, UK
vania@comp.leeds.ac.uk, scdd@leeds.ac.uk

Abstract: This paper describes development work in progress on tools to identify connections between researchers, as well as between researchers and business and other wider partners. The work is being carried out as part of a project, the Building Research and Innovation Networks (Brain) project, based at Coventry University in the UK with Leeds University as partners, and is part of a JISC funded Virtual Research Environment Programme. The Brain project aims to facilitate the building of Communities of Practice and networks of researchers and business and community partners to help enable the collective intelligence that potentially exists if these participants could be suitably engaged. In this endeavour, the project has explicitly identified the Research 2.0 approach as being central. Within the scope of this paper, only certain aspects of the project will be considered in any depth. The wider project includes work on business and knowledge related processes which impact on the nature and validity of the data used by tools such as those described here. Also central to the project is the building of Communities of Practice of researchers and other partners and the development of physical and virtual networks to support these. The development work on the tools described here was carried out by Dimoklis Despotakis and Ajdin Brandic.

Keywords: virtual research environment, brain project, research 2.0

1 Introduction

The tools discussed in this paper are being developed in response to ongoing user requirements identified by the project as well as conforming closely to the identified strategic institutional need to facilitate collaborative research focused around 8 themes. The techniques used can be considered part of those concerned with finding commonality between items and the tools discussed provide two main functions: Searching for researchers by keywords related to their work, and finding links from a

specified researcher to others. The key system components are the user input interface, a means of expanding keywords - using synonyms for example, the search mechanism, a means of filtering/weighting of results and the user output interface including suitable visualisation of information. The person linking tool adds an additional component which generates appropriate search keywords for an individual which can then be processed using the search functionality.

An example output for a person connection search is shown, illustrating the identification of expected close connections as well as ones from other disciplines. Also shown is an illustration of the use of the tools to create a map of linked topics and individuals around one of the broad strategic themes. Early evaluation of results shows a favourable reaction from users and considerable promise. Important future work seeks to extend the scope of coverage to wider research and business information, to using additional techniques including more adaptable and semantic methods and to looking in more depth at the underlying principles behind establishing connections, including applying pattern language-based approaches.

2 Requirements and Use Cases

The need for internet based services and tools to support communities of researchers has been generally recognised. Several national and international co-ordinating organisations and projects, such as JISC in the UK, Surfnet in the Netherlands and the EU Stellar Network have identified generic facilities and services which can facilitate collaborative research and complement discipline specific applications. Extensive user requirement analysis with researchers and other stakeholders at Coventry University has confirmed the need for certain functionality that correlates with requirements identified more widely. One such set of requirements relates to tools to support researchers finding potential collaborators or links to potential partners in business and the community. This arises in various forms in different stages of the research process. For example, at the inception stage of a possible piece of research, a typical need was expressed by a user as, "How do you find the people to talk to about an idea?"

At a later stage, when more detailed formulation of a research proposal or the writing of a paper is taking place, specific expertise, that could be outside the discipline of the main researcher or set of researchers, could be needed – perhaps in the area of data analysis or project evaluation. This would especially arise in cases of multi-disciplinary or inter-disciplinary research and in work that combined academic research with external activities in business or the community. A particular use case analysed by the Brain project illustrates the potential complexity of creating a suitable research team. This was a research call funded jointly by the Science and Social Science Research Councils in the UK on the theme of "Energy and Communities"1. This call involved subject areas ranging from environmental science, civil engineering and computer simulation through to psychology, sociology, economics and politics. A

[1] http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/Energy_and_Communities_Call_Specification_tcm6-34922.pdf

particular set of use cases has acted a central driver for the project and arose from the need to create collaborative networks to take forward work on 8 cross-disciplinary themes prioritised by the University as "Grand Challenges". A similar, but less clearly defined requirement arises when trying to identify groupings or clusters of researchers that may have the potential of working together or where the objective is to identify sub-disciplines within a larger area, but where the connecting themes are not known in advance. Examples of this which the Brain project has been engaged with concern finding connections between specific research groups and wider groupings of researchers for the purpose of the (UK) Research Evaluation Framework exercise that requires this for funding allocation purposes.

The basic methodology of the Brain project is to identify requirements, construct a structured model of the processes and services that could fulfill these requirements and then develop a prototype integrated environment based on this. These three parts of the work are closely intertwined and the Brain project adopts an Agile programming development methodology with short iterative cycles of development closely integrated with user requirements gathering, testing and evaluation. This paper describes some of the initial work carried out in the area of developing tools to identify connections, which forms part of the wider system to support collaborative research and innovation including discussion, networking and other services.

3 Techniques and Functional Components

It is not possible in this brief paper even to begin to indicate the considerable volume of research relevant to this area of work and this cursory introduction will only mention a few examples of work to set the wider context. Analysis of scientific and research networks and the connections between researchers that constitute them can reveal important characteristics and trends, such as the well known "six degrees of separation" property, described in one piece of research [1] as, "collaboration networks form "small worlds," in which randomly chosen pairs of scientists are typically separated by only a short path of intermediate acquaintances." Important examples of this work include an analysis of the Edmedia conferences [2] and an analysis of TEL Research Communities [3]. An extensive amount of software exists in the general field of social network analysis and is documented by organisations such as the International Network for Social Network Analysis[2].

The requirement considered here is about finding connections for the purposes described earlier and thus has a specific focus in comparison to the field in general. Numerous systems for finding experts exist, ranging from systems within individual organisations or particular membership networks, to those that aim to cover the web as a whole. A comparative evaluation of a number of these systems is made by Becerra-Fernandez [4]. Although a diverse variety of complex techniques are used by systems of this kind, it is possible to identify an underlying core set of functional elements used to implement them. One key generic component is to be able to

[2] http://www.insna.org/

identify what could be termed commonality – which could be between search terms and a document, between different researchers or researchers and businesses, and so on. This could be based on explicit or implicit characteristics. A search term has commonality with a document that contains it within the text – an explicit indication. Two researchers have commonality if they have read, cited or co-authored a particular paper – an implicit indication that arises from an aspect of their behaviour and which forms a central part of the analysis of the work mentioned above. Some quantifiable relative measure may be associated with this commonality. For example, two researchers who have referenced a paper would generally be considered to have a higher degree of commonality to two who have read the same paper, and two researchers who have co-authored a paper would be considered to have a still higher degree of commonality. A further simple metric may be the number of matches – the number of times a search term occurs in a document, the number of commonly referenced papers etc. This may need adjustment or normalisation in some form however – so that long and short documents or someone who has written a few papers can be compared with someone who has written many, for instance. In a more general sense this can include other features of adaptability that adjust the results to the characteristics of the data or the context. In some cases quantification of commonality or other analysis could be used to exclude certain results or weight them in some form that could be used in the visualisation – for instance grouping more strongly related items closer together.

More complex techniques that take into account indirect and secondary effects can sometimes be crucial to the success of this approach. A relatively simple and then more complex example will illustrate this. If we are searching for a keyword in a set of documents, we may wish to include synonyms or apply stemming techniques so that related terms are also searched for. This can be extended to include applying the concept of semantic distance [5], so terms which are more similar according to some criterion based on their place in some subject taxonomy, for instance, have a greater weighting. Perhaps the best known example of a more complex technique is the PageRank algorithm [6] used by Google and in various forms by other web search engines. The objective here is to quantify and thus rank the importance of a web page that contains a search term. The number of links to a page is used as the metric for this, but weighted by the number of links to those pages in turn and so on. Another well known technique will illustrate a different important aspect to using techniques of this kind. Recommendation engines used by businesses like Amazon base themselves on the commonality of customers reflected in the past purchases they have made to suggest new ones. An easily quantifiable success metric is available to the business in this case – what proportion of what customers actually buy are recommended items. Having a metric of some form like this is important to evaluate the success of techniques used and help choose and improve them. In the case of the research related examples considered, this will usually be more difficult and directly measurable metrics often not available. Nevertheless, having processes available to serve the same purpose - through user feedback and interaction for example – are still important, and incorporating these into the overall design is necessary to improve and evolve the systems implemented.

Although techniques such as these mentioned provide a powerful set of methods and a number are part of current project development activities, this paper focuses on

some of the initial development work which concentrated on the rapid creation of functional prototypes that could be deployed with users to meet real requirements. Although the techniques used for these were simple, they nevertheless provided usable functionality and allowed engagement of the project with researchers and others – a key priority at this stage, although improving and optimising the implementation is being carried out as a parallel process.

4 Implementation

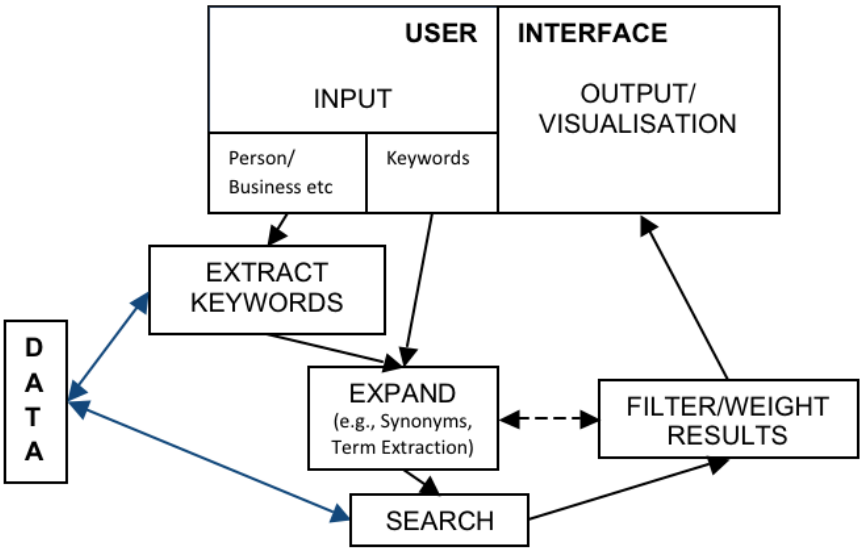

Fig. 1: Functional System Components

The diagram shows in outline form the basic functional components of the tool which incorporates in a simple form the key elements outlined in the previous section. For the initial prototype, these were implemented as follows.

4.1 User Input

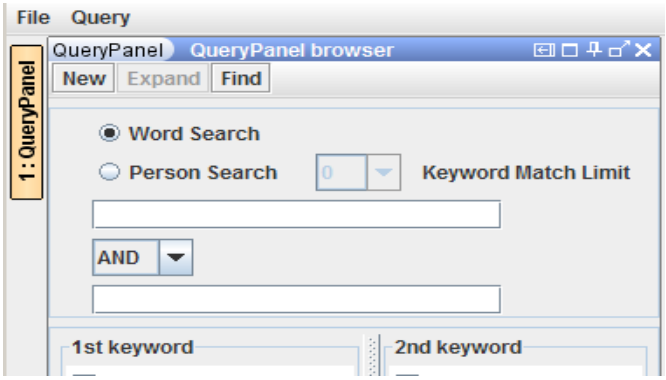

Fig. 2: User Input Interface

Shown above is the user input panel. Two facilities are provided in the prototype, searching by keyword or topic and grouping by person. For the keyword search, simple Boolean combination of terms is available.

4.2 Data

A central part of the wider Brain project was looking at the data relevant to research and the processes associated with it. In the general case, even for a particular requirement such as finding links between researchers, a very wide variety of data could be used in various ways. The project is interested in connections within the academic community as a whole as well as with wider business and community engagement. For the initial work however, data was restricted to the University, to provide a more limited scope for the requirement that could be evaluated more rigorously and then generalised appropriately. How this is being extended wider in developments currently taking place will be mentioned later. Data from a variety of sources has been used, providing information about researchers' expertise, interests, publications, projects etc. Linking information from these different data sets was done for the first time at the University by the project and proved a very considerable challenge. Some information was not available in an online form previously and special work had to be carried out to clean up and to link data where appropriate key fields did not exist. However, carrying out these tasks allowed valuable knowledge about research to be available for the first time, irrespective of the techniques used to process and analyse the information. The part of the project not detailed here relating to process has concentrated on how appropriate information can be made available in an up-to-date and reliable manner.

4.3 Commonality analysis

A brief indication of the range of functionality that could be used in this area was discussed earlier. Although some of these techniques are now being used in the ongoing development, the simpler techniques used in the initial prototype will be discussed here. The keyword search facility implemented was based on a simple string matching in the available data with the search words and selected synonyms, implementing simple Boolean combinations of these appropriately. Synonyms for the terms entered were generated using WordNet[3] and Disco[4] facilities and a checkbox facility provided for the user to choose these as desired.

For the person matching facility, the aim was to re-use a number of the components used in the keyword search. The keyword search process finds individuals whose associated information matches the keywords entered. Therefore, if appropriate keywords can be associated with an individual, an aggregation of the results from these as separate keyword searches can be used to determine the required person links. This raised the problem of how to generate these keywords. Where an explicit list of expertise areas was available in the data for an individual, for example, applying this approach would be trivial. However, applying this to other information was not as easy. Using the title of an academic paper as a keyword, for instance, would usually be very specific and therefore only usually match another academic if they were co-authors of the paper.

The synonym facility provided to expand the keyword search was not appropriate in this case and a different technique was used to implement this facility to generate keywords from sources of information such as the titles of papers. Among other techniques, two in particular using available web services were tried as part of this, the Yahoo Term Extraction service5 and the OpenCalais6 semantic metadata service . The Yahoo service proved to be more appropriate for use with publication titles especially and is the one used for the first prototype development, although the OpenCalais service is also being used in the system being developed.

Filtering/weighting results was looked at earlier as one of the components in determining commonality. In the early prototype system described here, adequate functionality for the keyword search could be provided without having to consider this area. However, for the person search this was an important consideration. In developing any system of this kind a balance has to be maintained between completeness and usability. When finding matches between people, a certain number of false positives can be expected. However, if these are too large as a proportion of results returned, the system will not be usable. Two techniques were used to tackle this problem. The first was the use of a stop list which filtered out certain words or phrases which were adjudged not to be useful in establishing connections, and was used after the stage of keyword expansion. For example, words like "research" and "university", are obviously too general to be of use. Considerable user testing and feedback was required in refining this stop list to limit matches to relevant ones, and

[3] http://wordnet.princeton.edu/
[4] http://www.linguatools.de/disco/disco_en.html
[5] http://developer.yahoo.com/search/content/V1/termExtraction.html
[6] http://www.opencalais.com/

the current list has over 1200 terms. The second technique used was to provide a user selectable filter parameter which would exclude terms which generated over a specified number of person matches. This allows searches to be run and then this parameter adjusted depending on the results.

4.4 Output and Visualisation

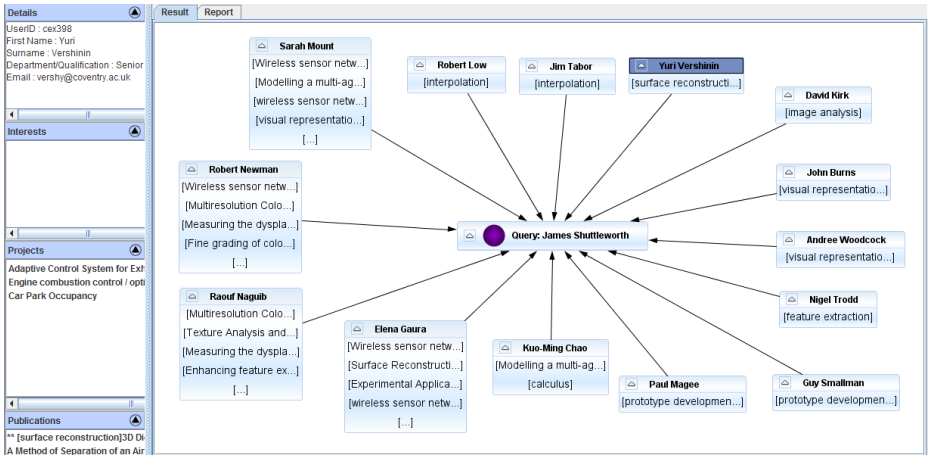

Fig. 3: Example Output

The output from a typical person connection search is shown above. Matched individuals have the items which were responsible for the connection displayed and highlighting an individual allows more detailed information about the match as well as other information about them to be shown in the side window. Individual matched items can be moved and hidden easily allowing particular features to be focused on if required. Multiple searches can be run and then tabbed between to allow different results to be compared and combined as necessary. A separate report view is also available which provides more detail about all the matches and which also provides the output in formats that can be exported into other applications for analysis and visualisation.

The example shown demonstrates one of the aspects of the system which allows new relationships and potential collaborations to be facilitated. In the illustration above, the researchers displayed on the left with a number of matches shown are members of research groups that the selected researcher, James, is part of. Thus their work (in the areas of Wireless Sensor Networks and Computer Analysis of Medical Images) can be expected to be already known to him. However, the tool has also picked up a variety of other researchers and associated research areas that in some cases are quite unexpected but nevertheless possibly relevant. These include mathematicians through the analytical techniques used by James, specialists in visual representation through the visualisation techniques he has used, and specialists in

different types of image analysis from disciplines as diverse as automotive engineering, shot peening, and Geographical Information Systems.

5 Use, Evaluation and further Development

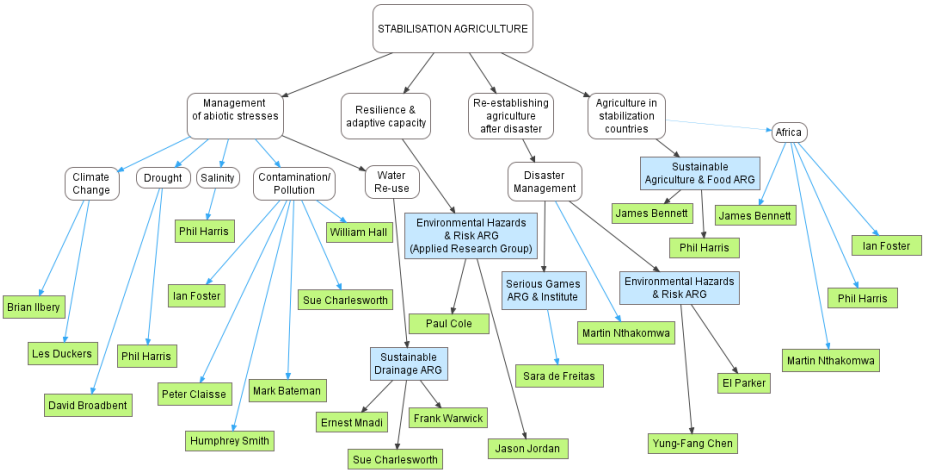

Fig. 3: Example Theme Mapping

As mentioned earlier, a key aim of the project was to be involved in fulfilling real requirements and solving real problems. Even in its relatively early stages the project and the tools it has created have had the opportunity to be embedded in key strategic initiatives and be tried out in practice. A significant amount of feedback and evaluation has been obtained working with individual researchers, research groups and research support staff, which has allowed significant iterative modification and improvement to be implemented, as well as further requirements that are currently being implemented to be identified. Space precludes detailed discussion of the many ways the tools developed have been used, but one example will be shown here. The diagram above shows a small section of one of the many visualisations of disciplinary and researcher links which the project working with appropriate researchers has generated, in this case for one of the "Grand Challenges" mentioned earlier - Sustainable Agriculture. Using both the keyword search and person link facilities together iteratively and the export facility mentioned earlier, powerful visualisations, in this case using the Visual Understanding Environment (VUE) application[7], can be constructed relatively easily. Used together with some of the other facilities that the project has made available, in the social networking field for instance, this provides a very significant capability to assist and facilitate collaborative research and innovation.

[7] http://vue.tufts.edu/

Key lessons learned from user engagement and feedback are summarised below, together with intended developments based on them and on the other aims of the project:

- Formal evaluation of the tools is mainly part of the forthcoming work of the project, but a co-evolutionary methodology utilising detailed feedback from users around specific use cases is the basic approach adopted. User response to the early prototypes has been very favourable in general. The system demonstrated its value from the first time it was used in practice by finding researchers for a particular initiative who were working on a common topic in different faculties but unaware of each other, and this has been repeated a number of times. As referred to earlier, this was partly a consequence of linking together information which had never been linked before as well as the expected and sometimes unexpected aspects of how the tools operate. In comparing previous attempts to manually carry out some of the tasks which the system has been used for it is also apparent that even semi-automated methods save a huge amount of time and make previously impossible analyses relatively trivial. The exercise has also helped to demonstrate the value of a more knowledge-based approach to university information and has fed directly into institutional strategic policy.
- Easy access to the tools and availability of current versions of software and up-to-date data are seen as a necessity, which in practice means implementing the tools as web-based applications. Currently the tools are implemented as a stand-alone PC application, mainly because the synonym generation facility used is only available in this form. Subsidiary web services for facilities like this will need to be developed if necessary.
- Extending coverage to include external information and being able to establish connections with researchers and others generally was both requested and a key aim of the project. This would require generalising how data is accessed and an implementation of the system which uses more general structured search, possibly implemented using Solr/Lucene, is part of the current development. Integrating information in RDF form together with the use of semantic search techniques is an intended further development. Because of the key aspect of the project relating to innovation as well as research, currently also being developed are ways to integrate business and other sources of information, using tools like OpenCalais and screen scraping and mashup tools as necessary. Considering commonality analysis in its more general sense could include facilities to recommend suitable papers to researchers, associate expertise and potential projects with funding etc. Because these requirements are linked, tools and services to deal with one can be used for others and the underlying knowledge set can be common, leading to the potential for a very powerful integrated environment.
- More powerful functionality to allow co-authorship and co-citations etc to be taken into account explicitly was seen as important and including a number of search, clustering and classification algorithms relevant to different contexts and types of information, is also necessary.
- Improvements in the visualisation algorithms and associated commonality techniques, for instance to reflect the strength of a connection by closeness, was a common request, as was the ability to manipulate and aggregate maps more easily, so

that multiple maps could be combined and connections linking to other connections generated automatically.

- Many improvements to the overall user interface and underlying functionality were suggested. Users often compared the tool to services like Google they were familiar with, requesting more flexible searching etc.
- Improvements to a number of auxiliary services used, such as the synonym mechanism, were requested. The current systems used, which are for a general audience, were considered too informal by some users. Work is being done to include more technical sources, thesauri and ontologies, such as the UKAT system8. Using systems of this kind allow more powerful commonality associations to be implemented, which have been found to be important to find less obvious connections – using measures of semantic distance for example.
- A considerable amount of feedback has been about the importance of including informal and tacit knowledge. Again, in considering "commonality" and how research topics and researchers link to each other, a number of assumptions have been made, such as that the closeness of match is the only criterion to be used. More sophisticated approaches are needed for understanding and representing information to take into account complementarity of knowledge and other considerations. A key part of the theoretical basis for the current project derives from earlier work carried out by members of the project team, in particular the Planet project9, which looked at how practice could be shared and represented – taking the use of Web 2.0 techniques in learning as an example, and the Connection project10 linked to this which looked at how connections between projects could be facilitated, particularly carrying this out for the set of projects that were part of the JISC Users and Innovation Programme (Emerge). A number of principles and techniques came out of this work, particularly involving the use of pattern language based approaches. The Brain project is seeking to further develop and extend some of these which are especially relevant to the tools discussed in this paper.

Acknowledgements

The authors wish to acknowledge the contribution of other members of the project team, Peter Haine, Derek Griffiths, Stella Kleanthous and John Tutchings. The contribution of the JISC in funding this work is also acknowledged.

[8] http://www.ukat.org.uk

[9] http://www.jisc.ac.uk/media/documents/programmes/usersandinnovation/planet%20final%20report.pdf

[10] http://cublogs.coventry.ac.uk/innovation/files/2010/08/jisc-connection-final-report.pdf

References

1. M.E.J.Newman (2001), The structure of scientific collaboration networks, *Proceedings of the National Academy of Sciences,* January 16, 2001 vol. 98 no. 2 404-409.
2. Ochoa, X., Mendez, G., Duval, E. (2009). Who we are: Analysis of 10 years of the ED-MEDIA Conference, ED-MEDIA 2009.
3. Marco Fisichella, Eelco Herder, Ivana Marenzi, Wolfgang Nejdl, (2010), Who are you working with? Visualizing TEL Research Communities, Retrieved on 5/7/2010 from: http://www.l3s.de/~herder/research/papers/2010/who_are_you_working_with.pdf
4. Irma Becerra-Fernandez, Searching for Experts on the Web: A Review of Contemporary Expertise Locator Systems, ACM Transactions on Internet Technology, Vol. 6, No. 4, November 2006, Pages 333–355.
5. Sowa, John F., (2000), Knowledge Representation, Logical, Philosophical and Computational Foundations.
6. Sergey Brin, Larry Page (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine, Proceedings of the 7th international conference on World Wide Web (WWW), Brisbane, Australia.