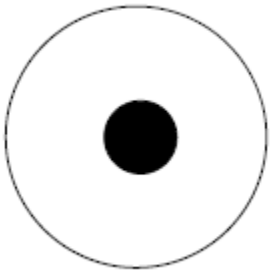
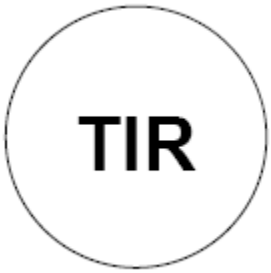




The IPSI BgD Transactions on Internet Research

Multi-, Inter-, and Trans-disciplinary Issues in Computer Science and Engineering

A publication of

IPSI Bgd Internet Research Society
New York, Frankfurt, Tokyo, Belgrade
January 2010 Volume 6 Number 1 (ISSN 1820-4503)

Table of Contents:

Pearls of Wisdom by an Advisor of European Commission:

Interview - The Essence of European Framework Projects
Pascall, Stephan 2

Belgrade University Research Efforts:

Document Filter Based on Extracted Concepts
Varga, Emil; Furlan, Bojan, Jakus, Grega; Milutinović, Veljko ……..... 5

Graph-Theoretic Formulation of Equations for Electrical Circuits with Mathematica
Tošić, Dejan 10

Visual PageRank: Improving the Random Surfer Model Using Visual Features
Diligenti, Michelangelo; Kovačević, Miloš 18

Applied Concept Modeling Techniques for Semantic Data Retrieval
Đorđević, Nemanja; Meštrović, Nenad; Mijović, Đorđe; Dimovski, Boris; Stančin, Sara 27

Hot Topics in Computer Science and Engineering
Tomažič, Sašo; Milutinović, Veljko 31

This is a special issue dedicated to: Bilateral Scientific Cooperation of Slovenia and Serbia, Project #BI-RS/08-09-007, in years 2008 and 2009 [E-speranto: vmesna platforma in razpis in shranjevanje večjezikovnih dokumentov s poudarkom na slovenščini, serbščini in ruščini]

The IPSI BgD Internet Research Society

The Internet Research Society is an association of people with professional interest in the field of the Internet. All members will receive this TRANSACTIONS upon payment of the annual Society membership fee of €100 plus an annual subscription fee of €1000 (air mail printed matters delivery).

Member copies of Transactions are for personal use only

IPSI BGD TRANSACTIONS ON INTERNET RESEARCH

www.internetjournals.net

STAFF	
Veljko Milutinovic, Editor-in-Chief	**Marko Novakovic, Journal Manager**
Department of Computer Engineering **IPSI BgD Internet Research Society** **University of Belgrade** **POB 35-54** **Belgrade, Serbia** **Tel: (381) 64-1389281**	**Department of Computer Engineering** **IPSI BgD Internet Research Society** **University of Belgrade** **POB 35-54** **Belgrade, Serbia** **Tel: (381) 64-2956756**
vm@etf.rs	**ipsi.journals@gmail.com**

EDITORIAL BOARD		
Lipkovski, Aleksandar	**Gonzalez, Victor**	**Victor Milligan, Charles**
The Faculty of Mathematics, Belgrade,	University of Oviedo, Gijon,	Sun Microsystems, Colorado
Serbia	Spain	USA
Blaisten-Barojas Estela	**Janicic Predrag**	**Milutinovic, Veljko**
George Mason University, Fairfax, Virginia	The Faculty of Mathematics, Belgrade	IPSI, Belgrade
USA	Serbia	Serbia
Crisp, Bob	**Jutla, Dawn**	**Neuhold, Erich**
University of Arkansas, Fayetteville, Arkansas	Sant Marry's University, Halifax	Research Studios Austria, Vienna
USA	Canada	Austria
Domenici, Andrea	**Karabeg, Dino**	**Piccardi, Massimo**
University of Pisa, Pisa	Oslo University, Oslo	Sydney University of Technology, Sydney
Italy	Norway	Australia
Flynn, Michael	**Kiong, Tan Kok**	**Radenkovic, Bozidar**
Stanford University, Palo Alto, California	National University of Singapore	Faculty of Organizational Sciences, Belgrade
USA	Singapore	Serbia
Fujii, Hironori	**Kovacevic, Branko**	**Rutledge, Chip**
Fujii Labs, M.I.T., Tokyo	School of Electrical Engineering, Belgrade	Purdue Discovery Park, Indiana
Japan	Serbia	USA
Ganascia, Jean-Luc	**Patricelli, Frederic**	**Takahashi, Ryuichi**
Paris University, Paris	ICTEK Worldwide L'Aquila	Hiroshima University, Hiroshima
France	Italy	Japan

Interview - ICT Projects in Serbia

Pascall, Stephan

1. How do you see the Serbian scientific capability in the field of ICT?

Republic of Serbia signed the Memoranda of Understanding and became associated to the 7FP on 13 June, 2007. By then, provisions for the association instruments had already been applicable since 1 January of that year. Serbia has been particularly successful in the food and agriculture, ICT and energy themes. Success rates so far in the ICT Calls 1-4 of FP7 are at 19%, thus above the 16% average of the Associated Countries (Albania, Bosnia and Herzegovina, Croatia, FYROM, Iceland, Israel, Liechtenstein, Montenegro, Norway, Serbia, Switzerland, and Turkey).

Nationally Serbia has defined a set of research priorities in ICT, namely in: a) Built-in electronic systems, b) Intelligent sensor-actuators, c) Management and control of complex distributed systems, d) Informatisation and digitalisation of libraries, e) Telecommunication systems for digital transmissions, f) radar and infrared identification and control systems, g) Expert systems, and h) Data safety.

Looking at the number of scientific publications in Serbia, the amount has steadily climbed from an average of just under 1000 per year in 2000-2003 to close to 2500 publications in 2008. So, we can see a steady progress.

2. How do you see the Serbian contribution to RTD in ICT in the West Balkans?

In general, regional cooperation offers many advantages because challenges are shared. Research cooperation in particular is facilitated by the fact that all Western Balkan countries are now associated to the 7FP. Serbia is gaining experience from taking part in the 7FP ICT-Theme and could become a good practice example.

In June 2006 the Steering Platform for Research Co-operation for the Balkans was launched. All Western Balkan countries are expected to be represented at Ministerial level and important work is being carried out addressing European, multilateral, and regional issues of Science and Technology policies. A further regional initiative, the WINS-ICT Dialogue Forum on ICT research for the Western Balkan Countries contributes to the work of the Steering Platform. Serbia hosted the first meeting of the Forum that took place in Belgrade in May 2009. The goal is to "foster possible harmonized solutions for common problems" in research related policy issues in ICT.

During the ICT Information and Brokerage event in Belgrade 20-21 May 2009, Ms. Jasna MATIĆ, Minister of Telecommunications and Information Society in Serbia, presented Serbia's strategy to create a sustainable economy and society through investments in science and technology.

The strategy involves establishing an innovation system engaging both public and private sector on a national, regional, and European level. Investments are foreseen in five key areas: 1) Human Resources, 2) Economic and Market Development, 3) Science Base, 4) Business R&D and Innovation and 5) Governance. To achieve this, the Serbian Ministry of Science and Technology has devised a series of actions, including a target for R&D Expenditures of 1% of GDP by 2014, €300m investments on key R&D infrastructure projects (funded through a joint loan from the European Investment Bank, the World Bank, and other international financial institutions and donations), and the identification of national research priorities.

3. What is the Serbian participation in the 7FP ICT-Theme?

In the first four ICT Calls there were 84 Serbian participants in submitted proposals and out of those 17 participated in retained proposals, giving the 19% average success rate mentioned earlier. The funding received amounts to just over €2.7m. In comparison, in FP6 the received funding totalled €24,600 (from participation in IST Call 6).

Dr. Stephan Pascall is with the European Commission, DG Information Society and Media.

4. What is DG INFSO doing to familiarise Serbian researchers with the possibilities offered by the 7FP ICT-Theme?

DG INFSO has together with the Serbian Ministries of Telecommunications and Information Society, and of Science and Technological Development organised a number of Information and Networking Events on the 7FP in general and the ICT-Theme in particular.

The events provide information about the current ICT Calls as well as technical, administrative, financial and legal aspect of project preparation and completion. The events also aim to foster project cooperation between Serbian research entities and potential partners from the EU27, Associated and Candidate Countries by offering networking opportunities.

More information and the details from a recent event in Belgrade can be found on http://fp7.mtid.gov.rs/

5. Why should Serbian researchers participate more in the 7FP ICT-Theme?

I think that the most important outcome of participation is to get access to a large pool of intellectual property and to become part of the network in a sector through the participation to the project, but more importantly also through the concertation meetings, etc. This will allow the organisations to build connections they can use for future cooperation, but also for business purposes.

Also, an obvious answer would be to take full advantage of the funding opportunities that are available. In total, the 7FP budget for 2007-2013 is €50.5bn. Under this umbrella the Cooperation Programme counts for €32.5bn, of which the ICT-Theme is the largest area with a budget of €9.1bn.

Apart from the clear financial benefits, there are catalytic effects and synergies that come with joint activities and cooperation with other European and international partners. Research is international in its nature and the 7FP holds opportunities in for example networking, sharing of expertise/facilities, development of new intellectual property and curriculum and staff training.

In a broader perspective, the Serbian participation in the Programme strengthens the European Research Area (ERA) and as such contributes to the Lisbon strategy objectives to make Europe the most competitive and knowledge based economy by 2010 and beyond.

6. How could they participate more?

As an associated country, the conditions for participation in consortia in FP7 are the same for Serbian stakeholders as for entities from the Member States. All proposals[1] are as a rule presented by multinational consortia of organisations/individuals from at least three different Member States or associated countries.

The ICT Proposers Day in Budapest 21-22 January saw the development of an online networking tool that provides support for finding partners and building consortia, http://ec.europa.eu/information_society/events/budapest_2009/index_en.htm

7. What kind of financial support is available?

In general, funding in the 7FP ICT-Theme is available through 3 funding schemes comprising 5 "instruments",

1. Collaborative Projects (CP)[2]
    a. Small or medium scale focused research actions (“STREP”)
    b. Large Scale Integrating Projects (“IP”)
2. Networks of Excellence (NoE)
3. Coordination and Support Actions (CSA)
    a. Coordinating or networking actions (“CA”)
    b. Support Actions (“SSA”)

The funding scheme, announced in the Calls for proposals, states: a) the purpose of the project, b) the type of activities involved, and c) the funding rate. Once specified in the Call, the instrument will not change and submitted proposals are evaluated against the same standards. In proposal preparations it is hence important to verify that your capabilities match the particular Call requirements with regards to funding scheme, along with other specified restrictions. Detailed information can be found in the "Practical guide to EU funding opportunities for research and innovation", ftp://ftp.cordis.europa.eu/pub/fp7/docs/practical-guide-rev2_en.pdf

[1] Except certain Support Actions
[2] Include SICA – Specific International Co-operation Actions

8. What is the participation of the EU15, EU12 and AC in the 7FP ICT-Theme?

The EU 12 and AC countries account for approximately 15% of the total project participations in the FP7 ICT-Theme Calls 1-4. Subsequently 85% of the participations come from the EU 15. Participant success rates are more on similar levels with the EU15 on 17%, the EU12 on 12% and the AC on 16%.

9. What is DG INFSO doing to help increase participation of the EU12 and AC in the 7FP ICT-Theme?

The EU wants to utilise the ICT R&D potential in all MS and AC to increase research excellence & competitiveness for the benefit of all. In this respect a series of actions are being taken including information and networking events, a technological audit and a top up call facilitating the inclusion of entities from the EU12 and AC into existing FP7 projects. A Working Group has been set up to identify the barriers to participation and measures to overcome them.

10. Where can I find information about the 7FP ICT-Theme?

- Participate in Information and Networking Events
- A supporting website of advice, information and documentation http://cordis.europa.eu/fp7/ict/
- Model grant agreements, guides, checklists, etc. http://cordis.europa.eu/fp7/find-doc_en.html
- Support services http://cordis.europa.eu/fp7/get-support_en.html
- An ICT Information desk for proposers' questions ict@ec.europa.eu
- An IPR Helpdesk http://www.ipr-helpdesk.org/index.html
- A Helpdesk for electronic proposal submission support@epss-fp7.org
- And a network of National Contact Points in Europe and beyond http://cordis.europa.eu/ist/ncps.htm

11. How can I find research partners for a project?

Methods to identify the right partners include networking at events as well as utilising online networks such as the ICT Proposers' Day 2009 website, the Ideal-ist project and CORDIS partner search (links below).

http://ec.europa.eu/information_society/events/budapest_2009/index_en.htm
http://www.ideal-ist.net/about-ideal-ist
http://cordis.europa.eu/fp7/partners_en.html

12. Can you offer any best practices in 7FP ICT-Theme participation?

Ask yourself…

- Do you clearly understand the pro's and con's of involvement in FP7?
- Are you willing to co-operate internationally?
- Do your capabilities match the Call requirements?
- Do you have resources to invest in time and money?
- Do you have a supportive team and organisation?
- Are there any favourable external research conditions to exploit?

What to do?

- Study the Workprogramme
- Set focussed, realistic objectives, and show how they will be achieved
- Consortium to include all stakeholders
- Good impact * good research = winner
- Write clearly for informed non-specialist
- Be within the scope of the call
- Chose the suitable programme
- Chose the suitable funding scheme
- Search for the right partners
- Keep the deadlines
- Be aware of the basic rules and guidelines
- Ask for help, look for local information days – look for the National Contact Points

What not to do?

- Address more than one objective
- Set broad, ambitious, unachievable objectives
- Include redundant partners
- Impact = a conference each year, a public website, and x published papers
- Buzzwords, empty statements

Stephan Pascall is a manager with extensive engineering and policy background and over 30 years' experience with the telecommunications and space industry and the European commission in a number of fields including telecommunications and its applications in space communications, Advanced mobile services, Trust and confidence services, Learning and culture, Regional development, environment and transport. Stephan Pascall is a recognized authority in Space Communications and an author of a number of books and publications in subject.

Document Filter Based on Extracted Concepts

Varga, Emil; Furlan, Bojan; Jakus, Grega; and Milutinović, Veljko, Fellow of the IEEE

Abstract*— *Document Filter Based on Extracted Concepts is integrating two Natural Language Processing (NLP) tools, ConceptNet and Antelope, to get the best results from their synergy in extracting a topic and important conceptions from a text. The proposed solution processes and extracts important information from specified text, and represents them in a form of a cloud (a fuzzy set) of concepts and contexts. Based on a selected concept, specific parts of a text would be magnified and thus displayed to the user as more significant. The interaction with the visually represented results is easy and intuitive.

Index Terms—, Concept Mining, Context Extraction, Fuzzy Logic, QA Systems, Concept Cloud, Natural Language Processing, Document Filtering.

1. INTRODUCTION

Recently, many systems were developed for information extraction from texts written in a natural language, based on: statistical and text mining approach (e.g. *Gate* [1], *Rapidminer* [2]) or natural language processing (e.g. *ConceptMine* [3], *Antelope* [4], and *ConceptNet* [5]).

Statistical approach is characterized primarily with greater accuracy for the larger processing corpus, easier implementation, and a long history of development and use [6] [7]. Nevertheless, if it is the case to process a shorter text, this approach does not give the best results. Well-known techniques for processing shorter texts are related to the field of computational linguistics [8] [9]. However, in processing general purpose texts, unrelated to a specific area the use of computational linguistics in most cases do not bring sufficient results [10].

Over time, the development of natural language processing tools with increasing computer performance and available memory made executing demanding techniques of artificial intelligence for PC programs possible.

Essentially, this system has been designed for extracting relevant concepts from some structural text unit (e.g. chapter, paragraph or sentence), where, based on a selected concept, specific parts of a text would be magnified and thus displayed to the user as more significant. Considering that these text units regarding their nature are not very long the decision was to involve NLP techniques in this project.

Discovered concepts are visually represented in a form of a concept cloud (*TagCloud* visualization [11] [12]). One of the benefits of this approach is that the more significant the concepts are, the bigger font size they have, which gives a more intuitive representation of relations between specific concepts and their importance in the text.

2. RELATED WORK

Question-answering systems (QA systems) are something that scientists have been continually developing since the beginning of the sixties. Early systems were made specifically for a certain domain like medicine or biology, where development was based on latest achievements in the fields of information extraction and information retrieval and a large part of papers have been published at the *Text REtrieval Conferences* (*TREC*).

Greater progress in a field of general topic QA systems started afterwards. One successful system was *Cymphony* [13] published in 1999 and it was based on the search for named entities and their relations (at the semantic level). At the University of Pisa, the development of a system called *Pisab* [14] in 2000 began and it was one of the first question answering systems based on concept indexing. Authors stated that the concept extraction technology was not sufficiently developed at the time to provide satisfactory results. At the same year, scientist from the University of Alicante in Spain worked on a QA system on a semantic level [15] based on *WordNet* lexicon [16], and it achieved better results. In the following years, more and more focus was shifted on the research of question answering systems based on a conceptual level.

Already in 2002, *AnswerBus* [17] provided responses using available web search engines and showed excellent results with approximately

Manuscript received May, 17, 2009.

Emil Varga, Bojan Furlan, and prof. Veljko Milutinović are with the School of Electrical Engineering, University of Belgrade, Serbia (e-mails: emil.varga@gmail.com, bojan.furlan@etf.bg.ac.rs, vm@etf.bg.ac.rs).

Grega Jakus is with the University of Ljubljana, Slovenia (e-mail: grega.jakus@fe.uni-lj.si).

seventy percent chance that the correct answer will be among the first five, while its general accuracy was about sixty percent. In the same year in which *Goose* [18] appeared, as one of the first systems that used *Open Mind Common Sense* [19], a common knowledge base that *ConceptNet* is based on. During the next years, the ideas about the influence of context on the QA systems' accuracy were proposed [20] [21].

Although there are various implementations of QA systems that employ different principles of concept extraction [22] [23], all of them mainly relay on an available knowledge represented in some form (e.g. text) and selection is made between results on statistical criteria. This is not the case with the presented problem where filtering is based on some small portions of text, usually paragraphs or sentences. Therefore, another approach is needed and the proposed solution is using, as its foundation, a synergy between *ConceptNet* and *WordNet*, that generally gives better results than when these two lexicons are used separately [24] [25].

3. PROPOSED SOLUTION

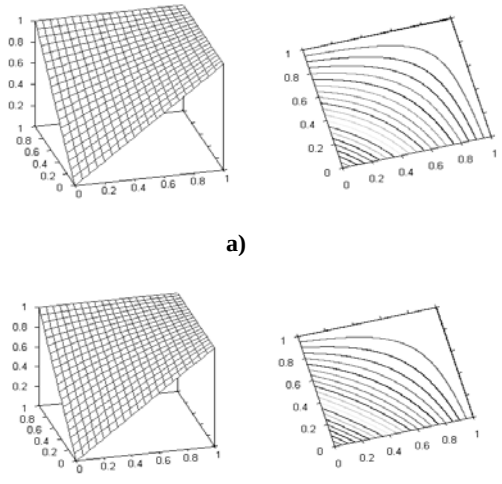
A Natural Language Processing framework, named *Antelope* (**A**dvanced **Nat**ural **L**anguage **O**bject--oriented **P**rocessing **E**nvironment) [4], is implemented in *.NET* technology and it is based on the *eXtended WordNet* [26] lexicon that extends the *WordNet* conceptualization. *Antelope* has a semantic parser and Word Sense Disambiguation (WSD) ability. It also integrates SUMO, formal high-level ontology [27] as well as the Stanford parser [28], statistical parser of great accuracy. On top of that, it includes tools for Named Entity Recognition (NER) and the detection of context, time and location. Because of these, Antelope was chosen as one of the tools used in this project.

ConceptNet is a very large semantic network of common knowledge that is created at the University of MIT. It is based on an aforementioned *Open Mind Commonsense Project* [19], where collected sentences from this project were coupled in a network of half-structured fragments of natural language representing concepts. This structure is a more flexible and suitable way of representing knowledge for NLP than formal logic [5]. *ConceptNet* also contains tools for text processing inherited from *MontyLingua* project [29], specifically for lexical, syntactic and semantic parsing.

Unlike lexicons and thesauri such as *WordNet*, *ConceptNet* in its knowledge base contains some ambiguities as they are in the nature of natural languages. Logically, this would be considered as incorrect, but as a semantic network of common knowledge, *ConceptNet* is more a representation of concepts from ā natural language perspective then from a logical perspective. This is simply inevitable as the semantic network is represented in fragments of natural languages. On the other hand, *ConceptNet* is in this way optimized for finding related concepts, while for ambiguity resolution it uses help from lexicons such as *WordNet*, *FrameNet* [30], *Longman's Dictionary of Contemporary English* - LDOCE and *Beth Levine's English Verb Classes*.

As mentioned before, *Antelope* in most cases recognizes named entities (e.g. human names, names of organizations, countries, cities, etc.) and some concepts, while in the experimental phase contains support for context recognition. How context can have its contribution in the recognition of terms and concepts, this information, although not always accurate, is stored in the system and used in the concept cloud. In addition, *Antelope* finds a lot less important concepts than *ConceptNet*, while *ConceptNet* gives good results in finding the existing concepts, as well as other related concepts, but it is not finding named entities and not recognizing the context.

Concerning that both tools have the possibility of concept extraction (although each in a different way), the resulting concept cloud is made as a union of produced concept sets. That is, if only one tool finds a concept, it enters the final set with its own weight. Otherwise, the weight is calculated as a fuzzy sum of given weights from both tools. Fuzzy logic is chosen as it appeared very useful for describing concept relations and development of semantic web [31], where the probabilistic T co-norm was used for summing (U (a, b) = a + b - a * b) because of its slower convergence (see *Figure 1*).



a)

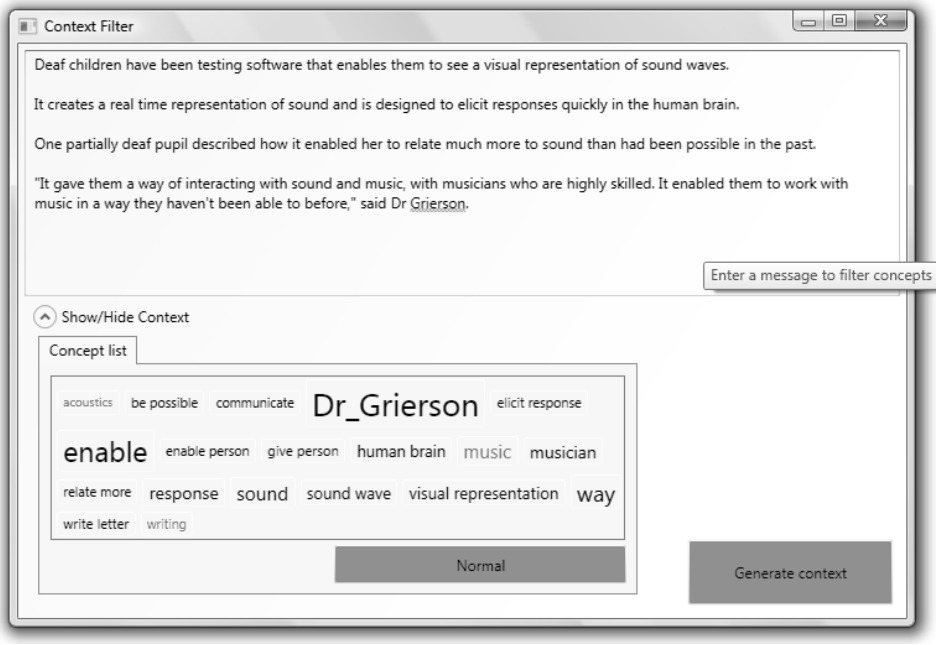
b)

Figure 1. a) Probabilistic T co-norm, b) Einstein's T co-norm

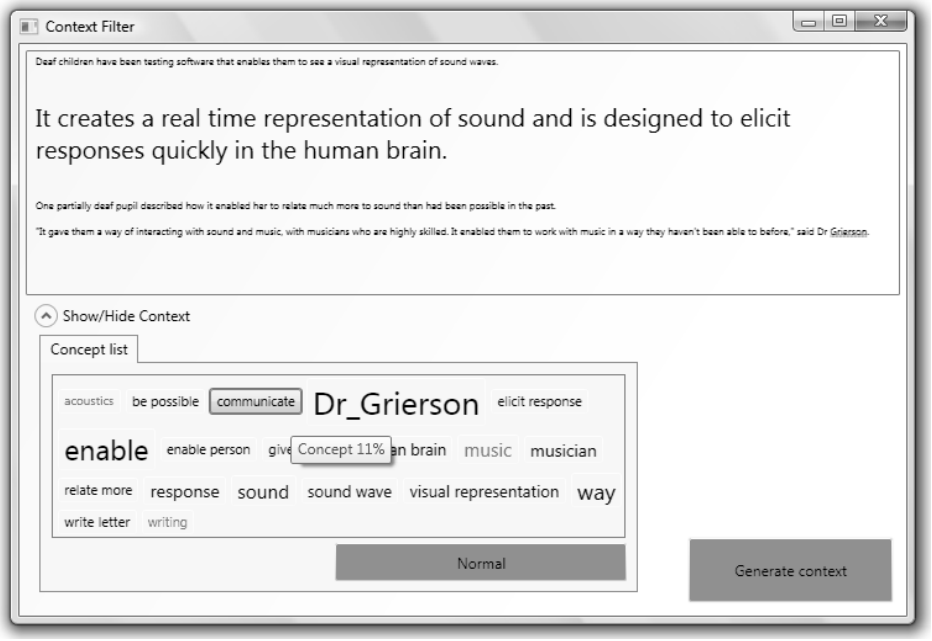
The extraction of concepts and context are generally followed by errors such as incorrectly discovered context in *Antelope*, or related concepts in *ConceptNet.* In order to reduce the number of errors during the extraction phase, a linear dependence was identified between the maximal weight of incorrectly detected concepts and the length of a question (more details in Section 5). There was no dependence found between the probability of incorrect context identification and any other given data, thus the system simply rejects context if its weight is not above the specified threshold.

In order to increase additionally the probability of *Antelope* finding the correct context, a simple fact is used: that it is harder to correctly conclude the topic from a bigger text then from a shorter one, consisting of most important concepts. For that reason, after all the results are gathered, the total list of available information is forwarded again to *Antelope* for context extraction, this time with a higher probability.

Other than that, a list of prohibited words that should not appear in the concept cloud is also introduced in order to further reduce the number of errors. Such words are usually articles, pronouns, numbers, interrogatives (like how, who, when, etc.) and other auxiliary words without much informative value.



a.) generated concepts cloud from the text,



b.) when *communicate* concept is selected.

Figure 2. Document Filter Prototype.

4. System Architecture

The system architecture is consisted of the NLP Engine module for the relevant information extraction and the presentation component that displays and interacts with the concept cloud. The communication between the NLP Engine module and the presentation component is asynchronous. The text that needs to be processed is delivered to the NLP module and after that, its internal thread continues the processing independently. When processing is finished, the module sends notification to the presentation component that the data is available. After that, the concept cloud is being created by converting the data into its graphic representation.

All concepts, contexts and named entities are composed of their text representation and the fuzzy value. These objects are elements of a fuzzy set that creates the concept cloud (see *Figure 2*). Concepts with their values form context in a wider sense, which together with the context generated from *Antelope*, produces a certain fingerprint of a processed text. This fingerprint is specific for each particular text unit and is similar for units with the same topic.

NLP engine module consists of the resources for connecting to *Antelope* and *ConceptNet*, filters for the words that should not appear in the cloud and two dedicated threads, one for a communication with the local *Antelope* resource, and other that has remote access to *ConceptNet*.

ConceptNet resource is implemented as a Web service in a *Microsoft Windows Communication Foundation* [32] technology in order to reduce the effect of the cold start delay, while *Antelope* is embedded in a client application, as its warm-up time is relatively short. This way parallel processing on both tools is gained, passing a part of a processing job to the client machine and avoiding heavy burden on the server. In addition, since *Antelope* is called twice, once to find available information from text and the other time, to extract the context from integrated results with *ConceptNet* more precisely, performances are improved accordingly, having in mind that calls from the local memory take up less time than remote server calls.

5. Tuning and Evaluation of Implemented Prototype

Both *ConceptNet* and *Antelope* impose certain errors into decision-making, as integrated NLP tools in the Document Filter Based on Extracted Concepts. Based on empirical results, in order to reduce the number of these errors, some design decisions were made. One was to ignore the results with values below the error threshold, then to ignore the results that contain more than two words and to ignore the results that are on the list of prohibited words.

The list of prohibited words, which contains all words that do not carry any important information, is introduced to filter out the words that are considered as invalid results. In addition, with the increase of the number of words in the result increases the chance that the result is incorrect. On a sample of about 50 randomly chosen texts it was noticed that a great majority (more then 93%) of results, which have more than three words and most of the results (above 87%) with more than two words, are incorrect. Therefore, the results that are considered to have a large likelihood of being erroneous are ignored in order to reduce the number of incorrect results.

The NLP tools as less likely mark results that have smaller values. Thus, context or concepts that have less value generally have a greater chance to be incorrect. This means that there is a threshold value for errors, so that if the results are below this value they have a high probability of being incorrect. To prove this claim, an empiric test of implemented prototype was carried out.

One of the basic project decisions in the test was to try to find dependence between the maximum value of incorrect results, or the error threshold, and the size of the text (more specifically the word count). The normal range for average structural units (paragraphs or sentences) are usually between ten and sixty words, therefore, tests were executed on texts with a word count that are in that scope.

The results showed that the number of incorrectly found contexts and the maximum value of an incorrect context are independent of the number of words in a text. However, the maximum value of incorrect concepts grows linearly with the increasing number of words in a text (see *Figure 4.*) according to the formula:

threshold for incorrect concepts = 7.45% + number of words in the text * 0.11%

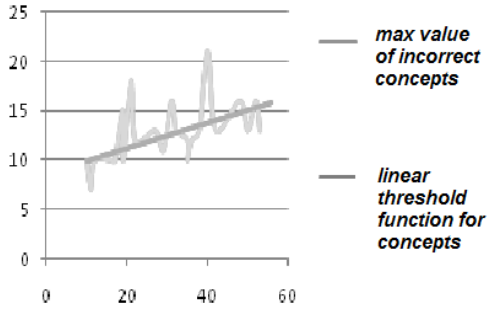


Figure 4. Linear approximation of the threshold function for incorrect concepts

Since there is always one up to four concepts that deviate from this formula, an attempt has been made to find a possible relationship. Particularly, the observed relationship was the one between the difference among the maximum value of incorrect concepts in a text and the limit value for error made by the previous formula, and a relative relationship between the highest value and the limit value for error. During the testing, no such existing relationship was found.

The already mentioned formula was selected for calculating the limit value for wrong concepts. A constant of 3% was chosen as the maximum value for incorrectly extracted contexts, since a high number of contexts found above this constant were correct and a high number of contexts found below were wrong.

6. CONCLUSION

The Document Filter Based on Extracted Concepts is conceived as a visual and easy to use application that employs synergy of two existing NLP technologies for the concept cloud extraction from text and filtering significant parts of a text, based on a selected concept. It is specialized for shorter texts, usually paragraphs or sentences, where it showed satisfying results.

7. ACKNOWLEDGMENT

Authors would like to thank Joe Sox for the excellent work on the *ConceptNetUtils* Library that he selflessly shared with all users of the Internet, professionals from Proxem that developed *Antelope* NLP tools, all researchers from MIT involved in development of *ConceptNet* and Radmila Radovanovic, Ana Varga, and Katarina Atanackovic for help in lecturing and advices from linguistics and semantic.

Authors are thankful also to Prof. Tomazic of U. of Ljubljana, for his help.

REFERENCES

[1] "GATE – General Architecture for Text Mining", University of Sheffield, Natural Language Processing Research Group, http://gate.ac.uk/, 25/6/2008

[2] "YALE – RapidMiner", http://rapid-i.com/content/blogcategory/38/69/, 25/6/2008

[3] ConceptMine, Scientio LLC, http://www.scientio.com/products/conceptmine/Pages/default.aspx, 25/6/2008

[4] Proxem Antelope 0.8, http://www.proxem.com, 25/6/2008

[5] H. Liu, P. Singh, "Commonsense reasoning in and over natural language," http://web.media.mit.edu/hugo/conceptnet/, 2004.

[6] H.P. Luhn, "The automatic creation of literature abstracts," IBM Journal, 1956.

[7] A. K. Heuttner, L. M. Schmandt, P. M. Andersen, P. J. Hayes, I. B. Nirenberg, "Automatic extraction in proceedings of the conference of the association for artificial intelligence," p.p 1089 - 1093, 1986.

[8] B. Sundheim, R. Grishman, "Message understanding conference - 6: A brief history," In Proceedings of the 16th International Conference on Computational Linguistics (COLING), pp. 466 - 471, 1996.

[9] "Message Understanding Conference MUC," http://www.itl.nist.gov/iaui/894.02/related_projects/muc/, 1997.

[10] S. Russel, P. Norvig., "Artificial Intelligence: A modern Approach," Prentice Hall Series in Artificial Intelligence, Englewood Cliffs, New Jersey, U.S.A 1995.

[11] D. Lemire i O. Kaser, "TagCloud Drawing: Algorithms for Cloud Visualization," 2007.

[12] P. Peroni, "On Tag Clouds, Metric, Tag Sets and Power Laws," http://blog.pietrosperoni.it/2005/05/25/tag-clouds-metric/, 2005.

[13] R. Srihari, W. Li., "Information Extraction Supported Question Answering. Cymphony Inc." TREC-8 conference, 1999.

[14] G. Attardi, C. Burrini, "The PISAB Question Answering System," Università di Pisa – Italy, TREC-9 conference, 2000.

[15] J.L. Vicedo, A. Ferrández, "A semantic approach to Question Answering systems," Universidad de Alicante, TREC-9 conference, 2000.

[16] G. A. Miller, "Wordnet: a lexical database for English langage," http://wordnet.princeton.edu/, 25/6/2008

[17] Z. Zheng, "AnswerBus - Question Answering System," University of Michigan, 2002.

[18] H. Liu, H. Lieberman, T. Selker, "GOOSE: A Goal-Oriented Search Engine With Commonsense," Adaptive Hypermedia and Adaptive Web-Based Systems, Second International Conference, AH Spain, 2002.

[19] "Open Mind Commons," http://commons.media.mit.edu/en/, 25/6/2008

[20] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, D. R. Karger, "The Role of Context in Question Answering Systems," 2003.

[21] E. D. Liddy, A. R. Diekema, O. Yilmazel, "Context-Based Question-Answering Evaluation," Syracuse University, New York, 2004.

[22] S. Kopparapu, A. Srivastava, P.V.S. Rao, "Minimal Parsing Key Concept Based Question Answering System," **Springer**, Lecture Notes in Computer Science, Volume 4552/2007, pp. 104-113, 2007.

[23] M. Nikravesh, "Concept-based search and questionnaire systems," **Springer-Berlag**, Soft Computing - A Fusion of Foundations, Methodologies and Applications, Volume 12, pp. 301 – 314, 2007.

[24] I. S. Eslick, "Searching for Commonsense," MIT, 2006.

[25] Ming-Hung Hsu, Ming-Feng Tsai, Hsin-Hsi Chen, "Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison," Japan, 2006.

[26] G. A. Miller, S. M. Harabagiu, D. I. Moldovan, "Wordnet 2 - a morphologicaly and semantically enchanced resource," http://xwn.hlt.utdallas.edu/, SIGLEX, 1999.

[27] "SUMO – Suggested Upper Merged Ontology," a candidate for IEEE standard upper ontology 1600.1, http://www.ontologyportal.org/, 25/6/2008.

[28] "The Stanford Parser 1.6," The Stanford Natural Language Processing Group, Stanford University. http://nlp.stanford.edu/software/lex-parser.shtml, 25/6/2008.

[29] H. Liu, "Montylingua a free, commonsense enriched natural language understander for English," http://web.media.mit.edu/hugo/montylingua/, 2004.

[30] C. J. Fillmore, C. F. Baker, J. B. Lowe, "The Berkley framenet project," http://framenet.icsi.berkeley.edu, 1998.

[31] E. Sanchez, F. van Harmelen, D. Dubois, J. Taylor, S. Thrun, "Fuzzy Logic and the Semantic Web," **Elsevier**, Capturing intelligence series, France, Université de la Mediterraneé, 2005.

[32] "Windows Communication Foundation WCF", .NET Framework Developer Center, http://msdn.microsoft.com/en-us/library/ms735119.aspx, 25/6/2008

[33] OMEROVIC, Sanida, TOMAZIC, Saso, MILOVANOVIC, M., TORRENTS, D. Knowledge processing and computer architecture. IPSI BGD Trans. Adv. Res., Jan. 2007, vol. 3, no. 1, str. 39-46

Graph-Theoretic Formulation of Equations for Electrical Circuits with *Mathematica*

Tošić, V., Dejan

Abstract—*This paper presents a new concept of teaching Electric Circuit Theory courses to engineering students, which uses a computer algebra system to formulate circuit equations. The paper focuses on a graph-theoretic formulation of equations, which completely specify an electrical circuit, by using Mathematica. The existing circuit simulators are mainly based on the matrix formulation of circuit equations, and are targeted for circuit design, but lack the pedagogical viewpoint. Original software CircuitSolver is presented, which is a set of carefully designed Mathematica functions that provide guidance in the process of solving an electrical circuit from the student's point of view. The circuits are assumed to be lumped, linear, and time invariant. The symbolic analysis is performed.*

Index Terms—Combinatorica, computer algebra system, symbolic analysis, circuit equation formulation, circuit graph theory

1. INTRODUCTION

IN engineering education it is considered important to learn Electric Circuit Theory (ECT). Grasping concepts and phenomena from circuit theory is important in understanding electric power systems, electronics, telecommunication and control theory, and microwave engineering. One of the important aims of a circuit theory educator is to help students acquire a "functional understanding" of the subject. In engineering education one implication of this is that students should learn to understand theories and models, and their relation to objects and events in the 'real' world, and learn to apply these models and theories. [1]

Education research regarding circuit theory has shown that the student's lack of qualitative understanding results in difficulties when solving quantitative problems. [2]

If educators wish to generate interest in ECT among the incoming engineering students, they should do a better job of promoting the profession by providing better teaching tools, and delivery methods, combining circuit theory courses with laboratory exercise. [3,4]

Mastering circuit analysis requires some degree of practice. One must grasp the idea of what, how, and even why something needs to be calculated, and yet be adept in algebraic manipulation. Often, the burden of algebraic manipulation causes the student to lose sight of the wood from the trees. [5]

New software tools that are available nowadays have dramatically influenced the possibilities of educational technology and play an important role in teaching Electric Circuit Theory. However, this does not simply suggest an improvement of hitherto used teaching methods: in many instances it is necessary to rebuild these methods from the ground up.

As reported in [6] the work of a teacher who chooses to stick with the classical methods of teaching and not to use the new software tools would necessarily be marked by many shortcomings: Students would be acquainted with outdated and less effective methods and procedures. Moreover, in the classic method of study a student must overcome the difficult barriers of mathematics, which makes the subject very unattractive.

The traditional educational process might

Manuscript received February 15, 2009. This work was supported in part by the Serbian Ministry of Science and Technological Development under Grant TR-11021.

D. V. Tošić is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, PO Box 35-54, 11120 Belgrade, Serbia (e-mail: tosic@etf.bg.ac.yu).

also imply issues that have only secondary importance for reaching the given objective. There are for example various alternative methods for finding solutions to concrete circuit problems that differ only in the laboriousness of calculations, but when using the computer these differences are negligible. Without computer-aided teaching, all this considerably diminishes teaching time that could otherwise be used for extending and intensifying the subject or for decreasing the number of lessons.

As observed in [7] students are usually somewhat confused by the ECT conventions used for defining the arbitrary signs of algebraic currents and voltages. Furthermore, they lack any methodology and do not have rigorous understanding of the concepts of node and loop. Thence, they often come up with wrong equations or an insufficient or redundant set of equations when they try to solve an electrical circuit.

Traditionally, computers were developed and used to expedite numerical calculations. A newer, and in the long run, very fruitful field is the manipulation of symbolic expressions. When these symbolic expressions represent mathematical entities, this field is generally called computer algebra. Today, with computer algebra systems (CAS), such as *Mathematica* [8], it is possible to calculate in minutes or hours the results that would (and did) years to accomplish by paper and pencil [9].

Symbolic analysis can provide many results which are simply not available from numeric simulation methods. Most importantly, they can provide explicit insight into the dominant behavior and properties of a circuit or system [10,11]. Therefore, CAS and symbolic techniques can help students acquire a "functional understanding" of Electric Circuit Theory and foster mastery of basic methodological skills in the study of electrical circuits.

This paper presents a new concept of teaching Electric Circuit Theory courses to engineering students, which uses a computer algebra system to formulate and solve circuit equations. Equation formulation for electrical circuits is essentially graph-theoretic in nature, so the presented concept makes extensive use of graphs to specify, represent and analyze circuits.

2. PROBLEM STATEMENT

This paper focuses on a graph-theoretic formulation of equations, which completely specify an electrical circuit, by using *Mathematica*. An original software package, *CircuitSolver*, has been developed in *Mathematica* to automate the equation formulation.

CircuitSolver is designed to be used as an instructional tool that is intended to help engineering students: (a) understand the conventions used for defining the arbitrary signs of algebraic currents and voltages, (b) learn how to proceed methodically in order to write a correct, complete and non-redundant set of circuit equations, and (c) acquire efficient skills in identifying circuit variables, equations, branches, nodes and loops.

The existing symbolic circuit simulators reported in the open literature [10] are primarily targeted to perform the symbolic simulations and are mainly based on the matrix formulation of circuit equations. They are good software tools for circuit design and evaluation of design alternatives, but lack the pedagogical viewpoint required for studying Electric Circuit Theory. However, *CircuitSolver* is a set of carefully designed *Mathematica* functions that provide guidance in a process of solving an electrical circuit from the student's point of view.

3. REPRESENTATION OF A CIRCUIT

Electric circuit analysis is based on formulation of the relevant equations and on their solutions. An example circuit, shown in Fig. 1, is considered to demonstrate how to set up the equations.

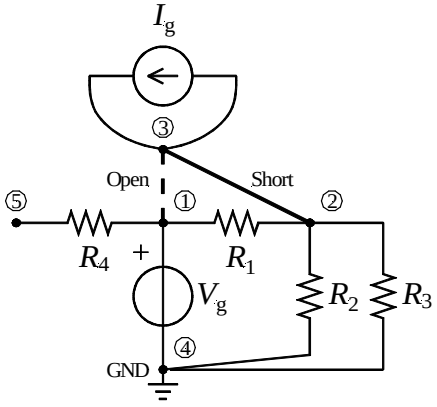


Figure 1. An example electric circuit.

The simplest way to specify the example circuit is to identify its nodes, for example, as a set **V = {1,2,3,4,5}**, and its elements, for example, as a set **E = { {{1,4}, "Vg"}, {{1,2},"R1"}, {{2,4},"R2"}, {{2,4}, "R3"}, {{1,3}, "open"}, {{3,2}, "short"}, {{1,5}, "R4"}, {{3,3}, "Ig"} }**. Each element is identified by an ordered pair of nodes and a unique label. The label is required to distinguish elements connected in parallel. This set of elements is often referred to as the netlist.

The circuit specification by the two sets **V** and **E** is the essence for the representation of the circuit by the directed graph **G = {V,E}**. In that case, the members of the set **V** are called vertices and the members of the set **E** are called edges. In the literature on graph theory, the terms "vertices" and "edges" are often used instead of "nodes" and "branches," respectively. If an edge is associated with the ordered pair **{j,k}**, then the edge is oriented from vertex **j** to vertex **k**. The graph **G** contains sufficient information for solving the circuit and it is the highest abstraction of the schematic shown in Fig. 1. However, from the educational viewpoint a pictorial representation of circuits is necessary.

Figure 2 shows another pictorial representation of the example circuit, which captures all structural characteristics of the circuit. Circuit element symbols are replaced by colored arrows and the element labels are represented as ***B*(T,L)** where **T** represents the type (**R**-Resitor, **V**-Voltage source, etc.) and **L** denotes a unique label. The arrows represent the assumed references for voltage and current, the positive voltage at the “upstream” end of the arrow and the current reference in the direction of the arrow. In general, if current is indicated as flowing from the “high” (+) voltage terminal to the “low” (–) voltage terminal of an element, the current conducted by the element and the voltage developed across the element to cause this flow of current are said to be in associated reference polarity.

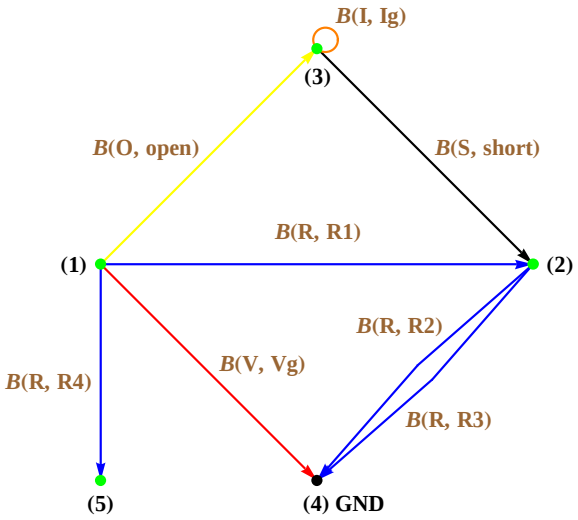


Figure 2. An abstract pictorial representation of the example circuit.

The representation of graphs takes on different requirements depending on whether the intended consumer is a person or a machine. Computers process graphs best as data structures such as adjacency matrices or lists. People prefer a visualization of the structure as a collection of points connected by lines, which implies adding geometric information to the graph. In this paper, the circuit graph is represented by the *Combinatorica* graph structure. This proves especially convenient when programming in *Mathematica* and provides the student with a flexible way to store embedding and drawing information associated with a graph. [12]

Here is the *Combinatorica*'s circuit graph representation:

```
circuit =
 Graph[{{{1, 4}, EdgeColor → Red, EdgeLabel → B["V", "Vg"],
    EdgeWeight → Vg, EdgeLabelPosition → UpperRight},
   {{1, 2}, EdgeColor → Blue, EdgeLabel → B["R", "R1"],
    EdgeWeight → R, EdgeLabelPosition → {0.1, 0.07}},
   {{2, 4}, EdgeColor → Blue, EdgeLabel → B["R", "R2"],
    EdgeWeight → 2 R, EdgeLabelPosition → {0.04, 0.1}},
   {{2, 4}, EdgeColor → Blue, EdgeLabel → B["R", "R3"],
    EdgeWeight → 2 R, EdgeLabelPosition → {0.15, -0.1}},
   {{1, 3}, EdgeColor → Yellow, EdgeLabel → B["O", "open"],
    EdgeWeight → 0, EdgeLabelPosition → UpperLeft},
   {{3, 2}, EdgeColor → Black, EdgeLabel → B["S", "short"],
    EdgeWeight → 0, EdgeLabelPosition → UpperRight},
   {{1, 5}, EdgeColor → Blue, EdgeLabel → B["R", "R4"],
    EdgeWeight → R / 2},
   {{3, 3}, EdgeColor → Orange, EdgeLabel → B["I", "Ig"],
    EdgeWeight → Ig, EdgeLabelPosition → {0.2, 0.07}}},
  {{{-1, 0}, VertexLabel → "(1)",
    VertexLabelPosition → {-0.03, 0}},
   {{1, 0}, VertexLabel → "(2)",
    VertexLabelPosition → {0.09, 0}},
   {{0, 1}, VertexLabel → "(3)",
    VertexLabelPosition → {0.033, -0.06}},
   {{0, -1}, VertexLabel → "(4) GND", VertexColor → Black,
    VertexLabelPosition → {0.15, -0.05}},
   {{-1, -1}, VertexLabel → "(5)",
    VertexLabelPosition → {0.03, -0.05}}},
  EdgeLabelColor → Brown, VertexColor → Green]
```

The name of the circuit graph structure is **`circuit`** and its header is **`Graph`**. The first element of this data structure is a list of edges; associated with each edge in the graph is a list whose first element is an ordered pair representing the edge. The second element of the structure is a list of vertices; associated with each vertex is a list whose first element is the position (coordinates) of the point representing the vertex.

Global information about how we want the graph drawn is stored as options in the graph data structure following the list of vertices. All local options that affect a specific vertex or edge are stored in the list corresponding to that particular vertex or edge.

The option **`EdgeColor`** sets color of objects representing edges, **`EdgeLabel`** sets edge labels, **`EdgeLabelColor`** sets color of edge label, **`EdgeLabelPosition`** sets position of edge label, **`EdgeWeight`** assigns weights to edges.

Edge labels are used to store element identification and edge weights are used to store element values.

The option **`VertexColor`** sets color of objects representing vertices, **`VertexLabel`** sets vertex labels, **`VertexLabelPosition`** sets position of vertex label.

Here is the circuit netlist extracted from the graph data structure by the *CircuitSolver* function **`CircuitNetlist`**:

```
CircuitNetlist[circuit]
```

```
The graph of the circuit is connected.

Self-loops exist.

Pendant (hanging) branches exist.
```

	Type, ID	+Node	-Node	Value	Description
1	B[V, Vg]	1	4	Vg	Voltage source
2	B[R, R1]	1	2	R	Resistor
3	B[R, R2]	2	4	2 R	Resistor
4	B[R, R3]	2	4	2 R	Resistor
5	B[O, open]	1	3	0	Open
6	B[S, short]	3	2	0	Short
7	B[R, R4]	1	5	$\frac{R}{2}$	Resistor
8	B[I, Ig]	3	3	Ig	Current source

The circuit graph and netlist clearly show the assumed reference directions to the student. The corresponding circuit variables, associated to the edges, are obtained by the *CircuitSolver* function **`CircuitVariables`**:

```
CircuitVariables[circuit, v, i]
```

$i_{\{1,2\}\mathrm{R1}}$ $v_{\{1,2\}\mathrm{R1}}$

$i_{\{1,3\}\mathrm{open}}$ $v_{\{1,3\}\mathrm{open}}$

$i_{\{1,4\}\mathrm{Vg}}$ $v_{\{1,4\}\mathrm{Vg}}$

$i_{\{1,5\}\mathrm{R4}}$ $v_{\{1,5\}\mathrm{R4}}$

$i_{\{2,4\}\mathrm{R2}}$ $v_{\{2,4\}\mathrm{R2}}$

$i_{\{2,4\}\mathrm{R3}}$ $v_{\{2,4\}\mathrm{R3}}$

$i_{\{3,2\}\mathrm{short}}$ $v_{\{3,2\}\mathrm{short}}$

$i_{\{3,3\}\mathrm{Ig}}$ $v_{\{3,3\}\mathrm{Ig}}$

The relation between the voltage and current in an element, known as the v-i relation, defines the element's characteristic. The element equations are generated by the *CircuitSolver* function **`CircuitElementEquations`**:

```
CircuitElementEquations[circuit, v, i]
```

$$v_{\{1,4\}\mathrm{Vg}} = V_{\mathrm{g}}$$
$$v_{\{1,2\}\mathrm{R1}} = R\, i_{\{1,2\}\mathrm{R1}}$$
$$v_{\{2,4\}\mathrm{R2}} = 2\, R\, i_{\{2,4\}\mathrm{R2}}$$
$$v_{\{2,4\}\mathrm{R3}} = 2\, R\, i_{\{2,4\}\mathrm{R3}}$$
$$i_{\{1,3\}\mathrm{open}} = 0$$
$$v_{\{3,2\}\mathrm{short}} = 0$$
$$v_{\{1,5\}\mathrm{R4}} = \frac{R\, i_{1,5\ \mathrm{R4}}}{2}$$
$$i_{\{3,3\}\mathrm{Ig}} = I_{\mathrm{g}}$$

4. Circuit Incidence Matrix

Consider a connected directed circuit graph with n vertices and m edges. The all-vertex incidence matrix $\mathbf{A}_{\mathrm{c}} = [a_{ij}]$ of the graph has n rows, one for each vertex, and m columns, one for each edge. The element a_{ij} is defined as follows: $a_{ij} = 1$, if the j-th edge is incident out of the i-th vertex; $a_{ij} = -1$, if the j-th edge is incident into the i-th vertex; $a_{ij} = 0$, if the j-th edge is not incident in the i-th vertex or if the edge is a self-loop.

An $(n-1)$ rowed submatrix of $\mathbf{A}_{\mathrm{c}}$ is referred to as an incidence matrix of the graph. The vertex that corresponds to the row that is not in $\mathbf{A}_{\mathrm{c}}$ is called the reference vertex of the incidence matrix. The incidence matrix is a coefficient matrix of Kirchhoff's current law. [13-16]

Assuming that vertex 4 is the reference vertex, the incidence matrix of the example circuit is computed by the *CircuitSolver* function **`CircuitIncidenceMatrix`**:

```
CircuitIncidenceMatrix[circuit, 4]
```

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}$$

The last column of the incidence matrix has only zero entries because it corresponds to the self-loop.

Here is the *Mathematica* code for computing the incidence matrix and it can be useful to show the computational details to the student:

```
CircuitIncidenceMatrix[g_Graph,
  n_Integer?Positive] :=
 Module[{e = Edges[g], t},
   t = Table[Which[SameQ @@ e[[j]], 0,
      First[e[[j]]] == i, +1,
      Last[e[[j]]] == i, -1, True, 0],
     {i, V[g]}, {j, M[g]}];
   Delete[t, n]] /; n ≤ V[g]
```

The above code is almost self-explanatory. The *Combinatorica* function **`Edges`** extracts the edges of a circuit graph ***g***, **`V`** gets the number of vertices, and **`M`** gets the number of edges. The *Mathematica* function **`Table`** creates a two dimensional array that represents the incidence matrix, **`Which`** performs the required tests to compute the matrix coefficients, **`SameQ`** returns true if an edge is a self-loop, **`First`** takes the first vertex of an edge, **`Last`** takes the second vertex of an edge, and **`Delete`** deletes the row corresponding to the reference vertex ***n***.

The circuit equations generated by Kirchhoff 's current law (KCL) state that at any instant, the algebraic sum of the currents (i) entering a node in a circuit is equal to zero. The KCL equations are formulated by the *CircuitSolver* function **`CircuitKCL`** (assuming that vertex 4 is the reference vertex and that i is the current variable symbol):

```
CircuitKCL[circuit, 4, i]
```

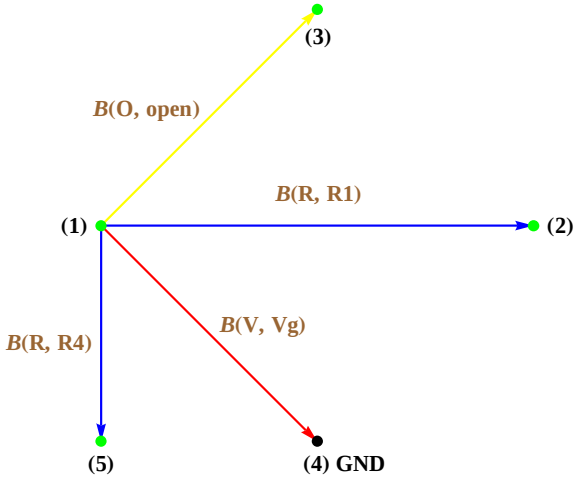
$$i_{\{1,3\}\mathrm{open}} + i_{\{1,2\}\mathrm{R1}} + i_{\{1,5\}\mathrm{R4}} + i_{\{1,4\}\mathrm{Vg}} = 0$$
$$-i_{\{1,2\}\mathrm{R1}} + i_{\{2,4\}\mathrm{R2}} + i_{\{2,4\}\mathrm{R3}} - i_{\{3,2\}\mathrm{short}} = 0$$
$$i_{\{3,2\}\mathrm{short}} - i_{\{1,3\}\mathrm{open}} = 0$$
$$-i_{\{1,5\}\mathrm{R4}} = 0$$

5. Circuit Loop Matrix

A path in a circuit graph is an ordered sequence of edges having the property that any two consecutive edges in the sequence share a common vertex. If the initial vertex is identical to the terminal vertex, then the corresponding path is called a loop (each vertex is traversed only once when tracing the loop). In other words, a loop is a connected regular subgraf of degree 2. [13-16]

Loops can be generated from the circuit tree that is obtained by the *CircuitSolver* function **`CircuitTree`** (the tree contains edge 1, i.e. the voltage source edge):

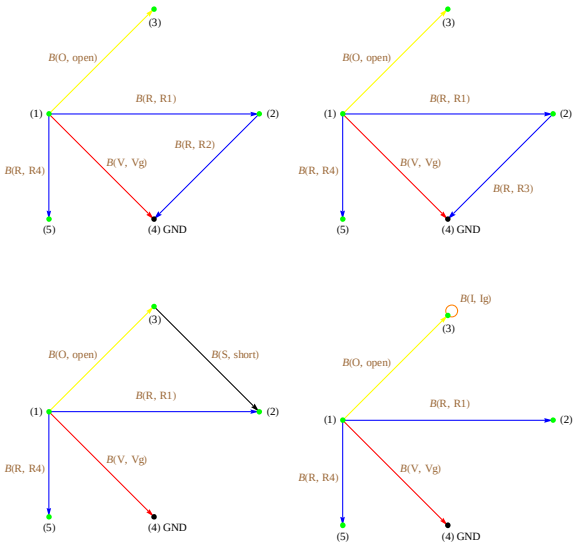
```
CircuitTree[circuit, 1] // ShowGraph
```



For a connected circuit graph, a tree is any subset of edges of the graph that connects all the vertices, but which contains no loops. The complement of a tree is called a cotree. Each edge contained in a tree is called a twig (or branch). Each edge contained in a cotree is called a link (or chord). If all links associated with a given circuit tree are removed, then the links replaced one at a time, it can be seen that each link defines a loop called a fundamental loop or f-loop (or fundamental cycle). [13-16]

`CircuitFundametalCycles`, which is a *CircuitSolver* function, constructs f-loops (for a tree that contains edge 1):

```
CircuitFundamentalCycles[circuit, 1]
```



The loop matrix $\mathbf{B}_c = [b_{ij}]$ of the circuit graph has m columns, one for each edge, and has one row for each loop in the graph. The element b_{ij} is defined as follows: $b_{ij} = 1$ if the j-th edge is in the i-th loop and its orientation agrees with the loop orientation; $b_{ij} = -1$ if the j-th edge is in the i-th loop and its orientation does not agree with the loop orientation; $b_{ij} = 0$ if the j-th edge is not in the i-th loop. The submatrix of $\mathbf{B}_c$ corresponding to the fundamental loops defined by the links of a circuit tree is called the fundamental loop (or cycle) matrix of the circuit graph with respect to the circuit tree. [13-16]

Assuming that edge 1 is in the circuit tree, the loop matrix of the example circuit can be computed by the *CircuitSolver* function **`CircuitFundamentalCycleMatrix`**:

```
CircuitFundamentalCycleMatrix[circuit, 1]
```

$$\begin{pmatrix} -1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The circuit equations generated by Kirchhoff 's voltage law (KVL) state that at any instant, the algebraic sum of the voltages (v) around a loop in a circuit is equal to zero. The KVL equations are formulated by the *CircuitSolver* function **`CircuitKVL`** (assuming that edge 1

is in the circuit tree and that v is the voltage variable symbol):

```
CircuitKVL[circuit, 1, v]
```

$$v_{\{1,2\}\mathrm{R1}} + v_{\{2,4\}\mathrm{R2}} - v_{\{1,4\}\mathrm{Vg}} = 0$$
$$v_{\{1,2\}\mathrm{R1}} + v_{\{2,4\}\mathrm{R3}} - v_{\{1,4\}\mathrm{Vg}} = 0$$
$$v_{\{1,3\}\mathrm{open}} - v_{\{1,2\}\mathrm{R1}} + v_{\{3,2\}\mathrm{short}} = 0$$
$$v_{\{3,3\}\mathrm{Ig}} = 0$$

6. CIRCUIT RESPONSE

Analysis of an electrical circuit involves the determination of voltages and currents in various elements, given the element values and their interconnections. In a linear circuit, the v-i relations of the circuit elements and the equations generated by the application of KCL at the nodes and of KVL for the loops generate a sufficient number of simultaneous linear equations that can be solved for unknown voltages and currents. [13-16]

For a connected directed circuit graph with n vertices and m edges *CircuitSolver* formulates $2m$ equations: m element equations (v-i relations), $(n-1)$ KCL equations based on the incidence matrix, and $(m-n+1)$ KVL equations based on the fundamental loop matrix.

Assuming that vertex 4 is the reference vertex, edge 1 is in the circuit tree, **v** is the voltage variable symbol, and **i** is the current variable symbol, the circuit response is computed by the *CircuitSolver* function **CircuitResponse**:

```
CircuitResponse[circuit, 4, 1, v, i]
```

$$i_{\{1,2\}\mathrm{R1}} \to \frac{V_\mathrm{g}}{2R} \qquad v_{\{1,2\}\mathrm{R1}} \to \frac{V_\mathrm{g}}{2}$$
$$i_{\{1,3\}\mathrm{open}} \to 0 \qquad v_{\{1,3\}\mathrm{open}} \to \frac{V_\mathrm{g}}{2}$$
$$i_{\{1,4\}\mathrm{Vg}} \to -\frac{V_\mathrm{g}}{2R} \qquad v_{\{1,4\}\mathrm{Vg}} \to V_\mathrm{g}$$
$$i_{\{1,5\}\mathrm{R4}} \to 0 \qquad v_{\{1,5\}\mathrm{R4}} \to 0$$
$$i_{\{2,4\}\mathrm{R2}} \to \frac{V_\mathrm{g}}{4R} \qquad v_{\{2,4\}\mathrm{R2}} \to \frac{V_\mathrm{g}}{2}$$
$$i_{\{2,4\}\mathrm{R3}} \to \frac{V_\mathrm{g}}{4R} \qquad v_{\{2,4\}\mathrm{R3}} \to \frac{V_\mathrm{g}}{2}$$
$$i_{\{3,2\}\mathrm{short}} \to 0 \qquad v_{\{3,2\}\mathrm{short}} \to 0$$
$$i_{\{3,3\}\mathrm{Ig}} \to I_\mathrm{g} \qquad v_{\{3,3\}\mathrm{Ig}} \to 0$$

CircuitResponse performs symbolic circuit analysis, which is the process of producing an expression that describes a certain behavioral aspect of the circuit with one, some, or all the circuit elements represented as symbols. In this particular example, all circuit parameters are kept as symbols or symbolic expressions. [16]

Hand analysis of circuits, especially if they contain many elements, can be tedious and prone to error. In such cases, computer-aided design (CAD) is necessary to speed up the analysis procedure and to obtain accurate results.

7. CONCLUSION

Mastering concepts and phenomena from electrical circuit theory is very important in understanding electrical power systems, electronics, telecommunication and control theory, and microwave engineering. If educators wish to generate interest in this field among the incoming engineering students, they should do a better job of promoting the profession by providing better teaching tools, and delivery methods. New software tools that are available nowadays have dramatically influenced the possibilities of educational technology and play an important role in teaching Electric Circuit Theory.

It has been observed that engineering students are usually somewhat confused by the circuit theory conventions used for defining the arbitrary signs of algebraic currents and voltages. In addition, they lack any methodology and do not have rigorous understanding of the concepts of node and loop, so they often come up with wrong equations or an insufficient or redundant set of equations when they try to solve an electrical circuit.

In this paper a new teaching concept has been presented, based on *Mathematica* – a computer algebra system, in which a circuit is solved pedagogically from the student's viewpoint. Original software *CircuitSolver* has been presented. *CircuitSolver* is targeted at students or educators seeking a computer assisted learning in mastering Electric Circuit Theory and clarifying its concepts and procedures.

REFERENCES

[1] A.-K. Carstensenb and J. Bernharda, "Critical aspects for learning in an electric circuit theory course – an example of applying learning theory and design-based educational research in developing engineering education," In *Proc. 1st International Conference on Research in Engineering Education (ICREE)*, Honolulu, USA, June 22–24, 2007, pp. CD-ROM.

[2] J. Bernharda and A.-K. Carstensenb, "Activity based education in electricity and circuit theory," Final report, Project no: 035/99, Linköping University, Norrköping, Sweden, 2004.

[3] Lj. Trajković, "Teaching Circuit Courses to Engineering Students," in *Proc. CAS Education Workshop*, Seattle, USA, May 22, 2008, pp. CD-ROM.

[4] Y. Tsividis, "Turning students on to circuits," *IEEE Solid-State Circuits Newsletter*, pp. 6–9, Jan. 2008.

[5] L. R. J. Costa, M. Honkala, and A. Lehtovuori, "Applying the Problem-Based Learning Approach to Teach Elementary Circuit Analysis," *IEEE Transactions on Education*, vol. 50, no. 1, pp. 41–48, Feb. 2007.

[6] D. Mayer and J. Kus, "Teaching Electric Circuit Theory with the Help of Tutorial Software," Global Journal of Engineering Education, vol. 2, no. 1, pp. 77–85, 1998.

[7] F. de Coulon, E. Forte, and J. M. Rivera, "KIRCHHOFF: An Educational Software for Learning the Basic Principles and Methodology in Electrical Circuits Modeling," *IEEE Transactions on Education*, vol. 36, no. 1, pp. 19–26, 1993.

[8] S. Wolfram, *The Mathematica book*, 5th Ed. Champaign, IL: Wolfram Media, 2003.

[9] M. Trott, *The Mathematica GuideBook for Symbolics*. New York, NY: Springer Science+Business Media, 2006.

[10] D. V. Tošić and M. D. Lutovac, "Symbolic Simulation of Engineering Systems," In *Proc. 4th IEEE European Conference on Circuits and Systems for Communications (ECCSC'08)*, Politehnica University, Bucharest, Romania, July 10-11, 2008, plenary lecture, pp. XIV-XXIII.

[11] D. V. Tošić and M. D. Lutovac, "Advances in Symbolic Simulation of Systems," *The IPSI BgD Transactions on Advanced Research*, vol. 3, no. 1, pp. 9–14, Jan. 2007.

[12] S. Skiena and S. Pemmaraju, *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*. New York: NY, Cambridge University Press, 2003.

[13] L. O. Chua, C. A. Desoer, and E. S. Kuh, *Linear and nonlinear circuits*. New York, NY: McGraw-Hill, 1987.

[14] D. E. Johnson, J. R, Johnson, J. L. Hilburn, and P. D. Scott, *Electric Circuit Analysis*, 3rd Ed. Upper Saddle River, NJ: Prentice Hall, 1997.

[15] W-K. Chen (Editor), *The Electrical Engineering Handbook*. Burlington, MA: Elsevier Academic Press, 2004.

[16] Wai-Kai Chen (Editor), *Circuit Analysis and Feedback Amplifier Theory*. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2006.

Visual PageRank: Improving the Random Surfer Model Using Visual Features

Diligenti, Michelangelo and Kovačević, A., Miloš

Abstract–Standard techniques for estimating web page relevance neglect the information provided by the visual layout of the page. The most popular link-based technique, PageRank, uses a model of a random surfer to estimate the probability that he visits a page at any given time. This probability is assumed to be proportional to the relevance of the page. PageRank considers all outgoing links equally probable to be followed by the random surfer. Web designers use visual features to assign different degrees of relevance to the links and help users in browsing and selecting the information they need. Discarding this information makes the random surfer assumption more unrealistic and could have a strong negative impact in the accuracy of the predicted relevance estimations.

The proposed approach is based on the layout analysis that assigns to each link in a page a set of visual features. This set of features describes the visual appearance of the link and the logical context in which it appears. We created a model based on the real user observations to represent how humans select links based on their visual appearance. Finally, the user model and the visual information that we made available for each link are merged to estimate the probability that each single link will be followed in any given page. The experimental results showed that the visual PageRank improves the accuracy of the relevance estimation.

Index Terms—PageRank, user model, *visual analyses*

1. INTRODUCTION

COMPUTING the relevance of web pages is of fundamental importance for search engines which have to select a small set of results among a huge number of candidates that are matching a typical query. The average number of search terms provided by a user of a search engine is in the range of 1-2 terms (queries with more than five terms are rare). For such short queries millions of results are the norm, and assessing page relevance becomes fundamental in order to provide useful results to the users. Therefore a good scoring function is commonly considered to be the keystone for a successful search engine.

PageRank [8] is the most popular algorithm to estimate the relevance of' web pages. It exploits only the link-based information to achieve its goal. In particular, PageRank uses a model of a random surfer to estimate the probability that the surfer is at any given time visiting a page. This probability is assumed to be proportional to the relevance of the page. While PageRank considers all outgoing links equally probable to be followed by the random surfer, this assumption is very unrealistic. Web designers use visual features to assign different degrees of relevance to the links and to help users in browsing and selecting the information they need. Clearly, a big link in the upper part of a page has more relevance than a small link in the bottom and it should be weighted accordingly. .

Since pages are designed for humans, the meaning of the raw data obtained from a HTML source could be better understood if we could know the spatial relations among text, images and other objects. The goal of the presented research was to better model a real human user using the visual features of a page. In the proposed solution we employ a novel approach based on the layout analysis (obtained via an implicit rendering of the page), that assigns to each link a set of visual features. This set of features describes both the visual appearance of the link and the logical context in which it appears.

In the proposed algorithm, the rendering step computes the coordinates and relative positions of the objects in a page. Then, each page is represented as a hierarchical structure called *Visual Adjacency Multigraph,* in which nodes represent simple HTML objects like text

Manuscript received May 4, 2009. (specify the date on which you submitted your paper for review.)

Diligenti. M. is with the Department of Informatics, University of Siena, Italy (e-mail: michi@dii.unisi.it).

Kovacevic. M. (corresponding) is with the Department of Management, Technology of Building and Informatics, Faculty of Civil Engineering, University of Belgrade, Serbia (e-mail: milos@grf.bg.ac.rs)

and images. Directed edges connecting the nodes of the multigraph represent spatial relations between the objects on the browser screen. The visual information contained in the multigraph allows defining heuristics for the recognition of common page entities such as vertical and horizontal link lists, titles and subtitles, and paragraphs of text.

Thus, visual analysis enables more accurate representation of the page contents, which splits the page into different logical portions. Moreover, at the end of the layout analysis step, each link is tagged with useful visual features like: size, color, position in the page, logical page portion to which it belongs (left or right menu, body, footer, header, etc.).

In order to be able to take advantage of the richer page (link) representation, it is needed to create a model of how real users select links based on their visual appearance. We decided to create such a model by recording the browsing behavior of real users. Finally, the user model and the visual information available for each link were merged to assign different weights to links, computed as the expected probability that the link will be followed by a user.

The outline of the paper is as follows: Section 2 describes the procedure of the visual layout analysis and defines heuristics for the recognition of logical groups in a web page. Section 3 presents a more general surfer model that is able to capture a more complex browsing behavior than the simple random surfer underlying PageRank. In Section 4, we present how visual information is embedded into the generalized PageRank model. Section 5 describes how the navigating behavior of users and their process of selecting links are modeled directly from users' observations. In Section 6, experimental results on a predefined datasets are shown. Finally, the conclusions are drawn in Section 7.

2. VISUAL LAYOUT ANALYSES

The main goals of the visual layout analysis are:

- Detecting the visual features of objects in a page. In particular, in this paper we will concentrate on links (with the corresponding anchor text), extracting the visual features as the color of the anchor text and its size. The size of a link is measured as the area on which the user can perform a click to follow it.
- Detecting the logical groups of objects in a page as they appear in a browser window, processing the corresponding textual file in an HTML format [7].

2.1 Logical groups of objects

Displayable objects are text, images, active-x controls, form controls and other. Each object is bounded with a boundary polygon of *n* vertices (where *n* is usually 4) and neighboring edges are always perpendicular. The logical groups of objects that can be taken into account strongly depend on the task at hand. For example, for the page relevance estimation problem that is addressed in this paper, a meaningful partitioning of a page could be to detect the portions of the page (and the corresponding links) that belong to:

- Body that is the central portion of the page where the most relevant information is presented. In some application like text classification, it could be possible to further subdivide the body into paragraphs.
- Vertical and horizontal menus (sequence of links). In particular, the vertical menus can be further subdivided into left and right menus, which have different semantic properties in a typical page.
- Header, the upper part of a page.
- Footer, the portion of a page at the bottom, which is often well visually distinct from the body and contains information which is non-essential to the rest of the page like: copyright information, information about the tool used to create the page or information about the author of the page, etc.

We emphasize that it is not possible to rely on HTML tags to determine whether a portion of a page belongs to a certain context. For example, some web designers could use the <.P> tag for labeling text and links that belong to a paragraph. But that assumption is highly inaccurate because other designers could use different tags to create the layout of a paragraph. An obvious example of misuse of tags is the usage of the <TABLE> one. Tables are commonly used (in 88% of the cases) to organize the layout of a page and the alignment of other objects, but not to organize tabular data [6]. It is not possible to rely on the proximity of text groups in a HTML source code because of previously mentioned <TABLE> tag and table nesting. Two text elements, which are close to each other in a source file, could be on the opposite sides of the screen. In order to simulate the visual human recognition the following problems have to be solved:

- Definition of an appropriate data structure that will reflect the positions of the objects and the spatial relations among them.
- Definition of a general recognition heuristics for each logical group of interest to be applied on the generated data structure.

The very first step in the visual layout analysis is to parse the page, and to construct the DOM tree that reflects the structure of the HTML source (nesting of container and inline tags/objects). A rendering procedure should be defined to calculate screen coordinates of the

objects in a DOM tree such as pure text, links and images. The rendering procedure should imitate the behavior of popular browsers. At the end of the first phase we obtain a coordinate tree in which the leaf nodes are represented as the coordinates of polygons embedding links, text boxes, and images. Each bounded polygon is represented as a set of four or more vertices defined with (x, y) screen coordinates (being the upper left corner of the screen the origin of the coordinate system). Details about the rendering process and the construction of the coordinate tree can be found in [7].

2.2 Visual Adjacency Multigraph

The recognition of logical groups of interest requires the understanding of spatial relations among the elementary constituents of a page. For example, when looking for vertical link lists it is needed to know if links are positioned in a vertical sequential row. Therefore it is needed to transform the coordinate tree from the rendering process into a more appropriate structure from which spatial relations among objects could be easily inferred. Before we define that structure more formally, we introduce four types of spatial relations among nodes of a coordinate tree. Let C be a coordinate tree and $p \in C$, $q \in C$ be two objects in the tree with bounded polygons C_p and C_q respectively. We introduce following definitions:

Definition 1: $p \leftarrow q$ (p is immediately left to q) if and only if (iff) the following statements are true:

1. $C_p \cap C_q = \varnothing$ (i.e. the polygons are not overlapping)
2. There exists at least one segment AB parallel to the x-axis connecting a point A $\in C_p$ and B $\in C_q$ (the intersection of the two projection of the polygons Over the v-axis is non zero).
3. There exists at least one segment AB for which rule 2 is valid such that for each $r \in C \setminus \{p, q\}$, AB $\cap C_r = \varnothing$ (the overlapping of the projections over the y-axis are not entirely covered by other objects).

Similarly,

Definition 2: $p \uparrow q$ (p is immediately before, or upper to q) iff the following statements are true:

1. $C_p \cap C_q = \varnothing$ (i.e. the polygons are not overlapping)
2. There exists at least one segment AB parallel to the y-axis connecting a point A $\in C_p$ and B $\in C_q$ (the intersection of the two projection of the polygons Over the x-axis is non zero).
3. There exists at least one segment AB for which rule 2 is valid such that for each $r \in C \setminus \{p, q\}$, AB $\cap C_r = \varnothing$ (the overlapping of the projections over the x-axis are not entirely covered by other objects).

The relations $\rightarrow$, (immediately right to) and $\downarrow$ (immediately after or down) can be defined symmetrically to the definitions above.

Definition 3: The *Visual Adjacency Multigraph* (VAM) of document d (represented as a set C of entries from a coordinate tree) is a directed multigraph with a set of nodes N and four sets of edges $E_{\leftarrow}$ $E_{\rightarrow}$ $E_{\uparrow}$ $E_{\downarrow}$. These sets have the following characteristics:

1. $N = |C|$ that is each object in the coordinate tree is associated to a node in the graph
2. $E_{\leftarrow} = \{(p,q) | p,q \in N: p \leftarrow q\}$. $E_{\leftarrow}$ contains all couples of objects p and q which are in $\leftarrow$ relation (all couples of objects which are horizontal immediate neighbors).
3. $E_{\rightarrow} = \{(p,q) | p,q \in N: p \rightarrow q\}$.. Symmetric definition to 2.
4. $E_{\uparrow} = \{(p,q) | p,q \in N: p \uparrow q\}$. $E_{\uparrow}$ contains all couples of objects p and q which are in p $\uparrow$ q relation (all couples of objects which are vertically immediate neighbors).
5. $E_{\downarrow} = \{(p,q) | p,q \in N: p \downarrow q\}$. Symmetric definition to 4.

For each couple (p, q) in $E_{\leftarrow}$, $E_{\rightarrow}$, $E_{\uparrow}$ or $E_{\downarrow}$, the VAM also stores the type of objects that are connected (links, text, images, etc.), their distance (in pixels) and a Boolean value describing if the polygons C_p and C_q containing the objects are vertically aligned for $E_{\uparrow}$ and $E_{\downarrow}$, or horizontally aligned for $E_{\leftarrow}$ and $E_{\rightarrow}$.

Figure 1 shows the whole process in which a web page is transformed into corresponding VAM.

2.3 Recognition of logical areas in a web page

Given a page p and its VAM(p), we want to label the links that belong to vertical link lists (and more specifically to left and right menus), horizontal link lists, the body of the page, the footer and the header of the page. Unfortunately, designers have different approaches to organize the same information to be displayed on the screen. Each designer can use different HTML tags or apply various cascading style sheets. Therefore we cannot directly rely on the information contained in the tags, except for rendering the page. After the rendering has completed, the positions, the spatial characteristics (like size, colors), and the relative spatial relations among HTML objects contained in the VAM are known.

In the proposed system a set of heuristics is used to discover the logical partitions from the VAM representing a page. As an example of

heuristic used in the system, we describe the heuristic applied to detect vertical menus.

Vertical Menu Heuristic: a vertical menu is detected using VAM as a path through the nodes in $E_{\uparrow}$ that are associated to links, that are aligned and that are closer than a given threshold.

It is clear that, given a VAM, it is easy to create other heuristics for detecting different partitions that could be of interest for other applications. Refer to [7] for a more extended description of these heuristics.

A big advantage of the approach is its flexibility and generality: having information about relative relations among objects, no matter where they are positioned, it is possible to easily create new heuristics for extending the recognition to other logical groups. This allows the application of the same framework to other problems like: page classification, focus crawling, recommending systems, page segmentation, etc. Probably the weakest point of the approach is the complexity of the multigraph construction process. In the current implementation the entries of the coordinate tree are processed, and for each new entry we check all previously encountered ones for spatial relations. Therefore, the complexity is $O(n^2)$, where n is the number of entries in the related coordinate tree (in a typical page less than 100 entries are included in the tree). There is a lot of space to optimize the implementation of rendering and recognition tasks, but that was left for future work.

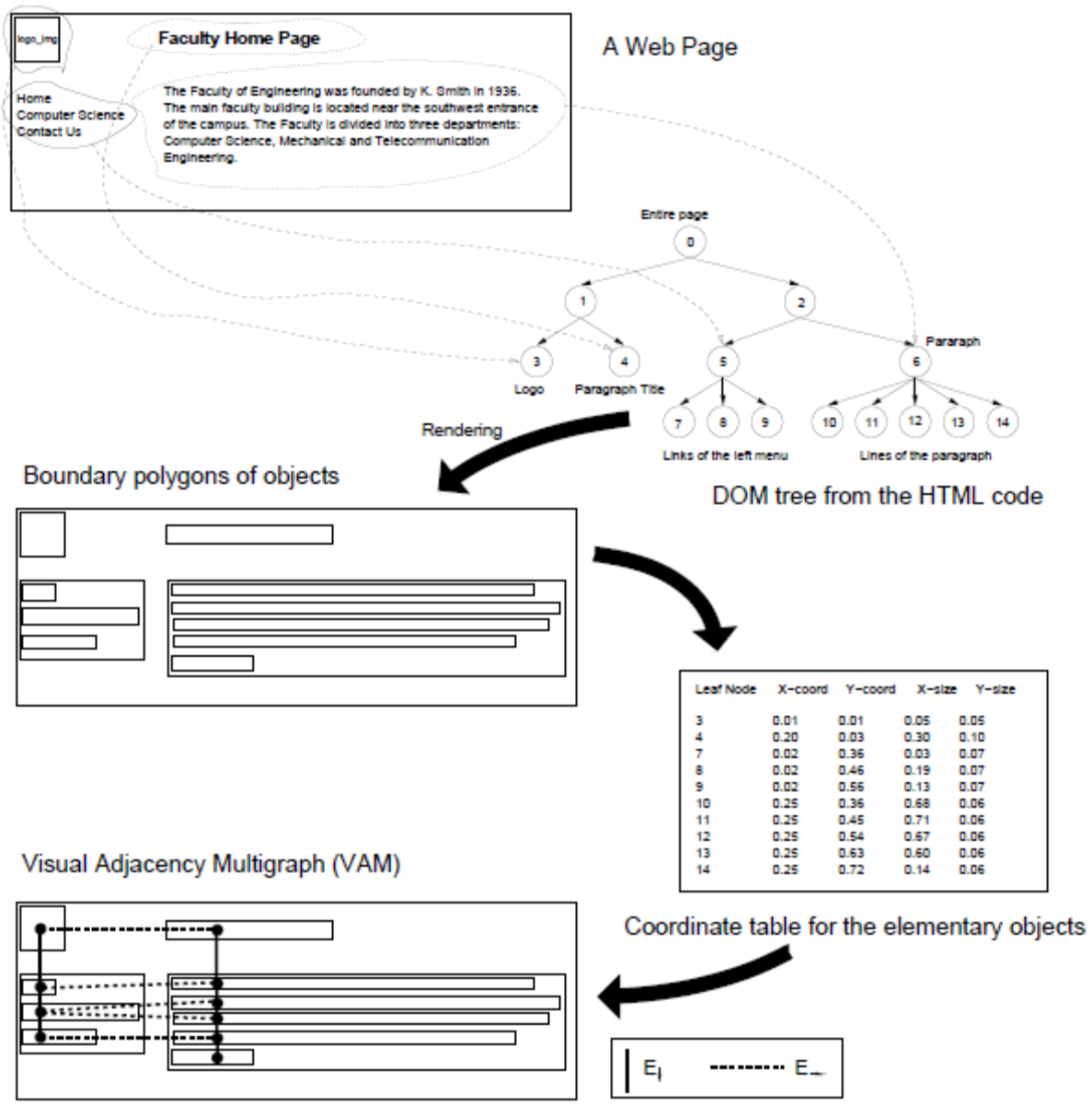


***Figure 1*: Given a web page, we render the page starting from the DOM tree extracted from the HTML source code. The output of the rendering module is the set of coordinates of the boundaries of the polygons containing the objects in the page. The Visual Adjacency Multigraph (VAM) is then extracted from the boundary polygons. Each edge of the VAM also includes information about the distance between the connected objects and if they are aligned. For sake of readability, this**

information is not displayed in the figure.

In [1] it is observed that a common surfer expects to see certain kinds of objects such as menus and banners in predefined screen areas of a page. This could allow sorting the elementary objects and restricting the areas of the screen in which to search specific logical objects resulting in a dramatic speed up of the process.

3. EXTENDED RANDOM SURFER MODEL

PageRank introduces a notion of page authority which is completely independent on the page content, whereas the authority only emerges from the topological structure of the Web. In PageRank, the authority is similar to the notion of citation in the scientific literature. In particular, the authority of a page *p* depends on the number of incoming hyperlinks (number of citations) and on the authority of the page *q* which cites *p* with a forward link. Moreover, selective citations from *q* to *p* are assumed to provide more contribution to the score of *p* than uniform citations. Hence, the PageRank x_p of *p* is computed by taking into account the set of pages *pa*[*p*] pointing to *p*. According to [8], the PageRank is computed using the following set of recursive linear equations:

$$x_p = \lim_{t\to\infty} x_p(t) = d \sum_{q\in pa[p]} \frac{x_q(t-1)}{h_q} + \frac{1-d}{|W|} \qquad (1)$$

$$\forall p \in W$$

Here, *W* is the set of pages in the considered Web, and $d \in (0,1)$ is a *damping factor* and h_q is the *hubness* of *q*, that is the number of hyperlinks contained in *q*.

The PageRank models a random walk on the Web graph of a surfer allowed to perform only two actions at each time step: the surfer jumps to a new random page with probability 1 - *d* or she/he follows one link from the current page with probability *d*. All these values are considered to be independent on the page *p*. Given that a jump is taken, its target is selected using a uniform probability distribution over all the I*W*I Web pages. Finally, the probability x(*p* I *q*, *I*) of following the hyperlink from page *q* to page *p* does not depend on the page *p* i.e. x(*p* I *q*, *I*) = α_q In order to meet the normalization constraint, α_q=1/h_q where the hubness of page *q*, h_q = *ch*(*q*), is the number of links exiting from page *q* (the number of children of the node q in *W*). This requirement cannot be met by sink pages, i.e. the pages which do not contain any link to other pages. In order to keep the probabilistic interpretation of PageRank, all sink nodes must be removed and their scores computed only after the model has converged. Typically, the initial page scores $x_p(0)$ are uniformly set to 1/I*W*I. Using well studied results of Markov chains theory [9], it can be easily shown that equation (1) converges. At convergence, the PageRank x_p of page *p* represents the probability that the random surfer will be located at *p* at any time step.

4. VISUAL PAGERANK

The main limitation of PageRank arises from the assumptions made about the underlying random surfer. Such assumptions are unrealistic and do not accurately model how a real surfer browses the Internet. In particular, it is strongly unrealistic that a user uniformly jumps to a random page when it decides that the current page does not lead to any further useful information. It is more likely that he will restart his surfing from some authoritative page or search engine. In order to improve the random surfer scheme, it is possible to use a biased distribution of the pages where the surfer will land after a jump [3].

Moreover, the surfer model assumes that all out-links of a page are equally likely to be followed. This is clearly unrealistic since users are very good in selecting specific links that they think to be useful. The surfer model could be easily extended to take into account a bias on the out-link probability [3, 4]:

$$x_p(t) = d \sum_{q\in pa[p]} x(p \mid q, I) x_q(t-1) + \frac{1-d}{|W|} \qquad (2)$$

$$\forall p \in W$$

However, whereas the model can be easily extended, it is not clear which semantic should be assigned to the link weights x(*p* I *q*, *I*).

In [3], this extended PageRank surfer model is used to create a topic-specific (or vertical) page rank, where the link strength *x*(*p* I *q*, *I*) is assigned to be proportional to the score assigned to the target page *p* by a classifier-by-topic. However, it is not clear how to bias the weights for non-vertical search.

Following the assumption that modeling a more intelligent and "real" surfer leads to a more accurate ranking of the pages [4], we propose to use the visual information to emulate the complex analysis that a user pursuits when looking at a document displayed on the screen. Since links have different visual characteristics, a user is influenced during the selection process by the link size, its color, position in the page, by whether the link is an image, and by many more factors. Using the visual approach, we aim at estimating the probability that a real user, looking at the page, would click on each single link on it.

5. ESTIMATING THE PARAMETERS OF USER MODEL

Since we aim at modeling the user behavior, we first tried to understand how real users are influenced by visual features and, then, to estimate model parameters directly from these observations. Let the links be subdivided into the following categories:

- Link with images: only reasonably big images with size greater than 1% of the screen were considered in this context.
- Links with emphasized anchor text. This category includes links with bold, capitalized or underlined anchor text or anchor text with font size at least one point bigger than the average text in the page
- Links with non-emphasized anchor text. These links do not "visually" emerge from the rest of the page.
- Other links.

Moreover, links are marked as belonging to the header, footer, body, left: and right menu of the page as described in section 2. Clearly, the above classification discards many other possible link categories. This simplification was needed to keep the number of the parameters of the user model to a reasonable level. We observed a set of 21 Internet users during their common search and browsing activities on the Web and we recorded their clicks. In particular, each user was recorded during 5 different information searching sessions for a total of 63 sessions and more than 2500 clicks.

	Header	Footer	Body	Right Menu	Left Menu
% clicks on links from	19.5	10.5	28.4	10.8	30.8

(a)

	Images	Emphasized Anchor Text	Standard Anchor Text
% clicks on links of type	23.8	13.0	6.1

(b)

***Table 1*: (a) percentage of clicks performed by the user on the links in each considered partition of a page. (b) for each link type, we recorded the percentage of elements that received at least one click for one user.**

	Header	Footer	Body	Right Menu	Left Menu
Link Strength	+0.06	-0.124	+0.101	-0.105	+0.122

(a)

	Images	Emphasized Anchor Text	Standard Anchor Text
Link Strength	+0.124	+0.09	-0.0248

(b)

***Table 2*: (a) visual strength of a link in each page partition. The strength is measured as the deviation of the observed user click probability from the baseline (the case of' a random surfer who randomly selects links without looking at the page). (b) The computed visual strength of each link category.**

For each visited page, we recorded the numbers of clicks on links from each considered category. In Table 1, there are reported the percentage of user clicks on a specific portion of the page and the percentage of clicked elements for each link category, respectively. The tables show that users tend on average to click more on image links and on links with emphasized anchor text. As expected, users click more likely on links in the body or left menus of a page. On the other hand, the user is less likely to click on links that are associated to standard anchor text (which is not visually emerging) and on links located in the footer and right menu of the page. These probabilities can not be directly used as weights of the user model since user behavior depends on the actual page, i.e. the user will not click on a link in the left menu if such menu is not present in the page. Similarly, a user could be very likely to click on links with emphasized anchor text simply because such links are more commonly found in a given Web document. As a conclusion, simply counting how many times the users click on a link of a specific category is not a good guess to measure how much a link category attracts the users. To solve this problem, the weights for a link of a specific category must be adapted to the average number of links of that category that are present on the Web pages. In order to do that, we define the visual strength S_t of a link type t as:

$$S_t = \frac{c_t}{C} - \frac{n_t}{N}$$

where c_t is the number of clicks on links of category t recorded during all sessions, C is the overall number of clicks recorded during all sessions, n_t is the number of links belonging to category t that appeared in the pages accessed during all sessions, and N the total number of links in all displayed pages. S_t represents the visual strength of each link type, measured as the deviation of the observed click probability (posterior probability) from the expected a-priori probability of that link category.

Obviously, a value S_t grater (or smaller) than zero means that specific link category is more (or less) likely to be clicked than expected when using a uniform distribution over the links. In general, the higher the value of S_t, the more likely the link will be selected by the user with a click and, therefore, the more bias should be assigned to the link during the random walk. Table 2 shows visual strengths for each considered link category.

We considered the bias due to the link type to be independent from the bias due to the position of the link in the page. This "naive" assumption is clearly not true, but it was needed to limit the number of parameters that had to be estimated while recording and analyzing the users' behavior.

For each page q each outgoing link to a target page p is classified according to its type (image, emphasized anchor text, standard anchor text) and its association to a given logical group (body, left/right menus, footer, header). Then the weight $x(p \mid q, l)$ of the link from page q to page p (the probability of the surfer to follow a link to page p given that he is located in q and he will not jump) is computed as:

$$x(p \mid q,l) = \frac{1}{h_q} + S^1_{q\to p} + S^2_{q\to p}$$

where $S^1_{q\to p}$ and $S^2_{q\to p}$ are the strengths due to link category and link position for the considered link, respectively. When $S^1_{q\to p}$ and $S^2_{q\to p}$ are equal to zero (i.e. there is no evidence that the visual features affect the user behavior), equation (2) becomes the standard PageRank equation. These parameters are then normalized for each page q to sum up to one and are then inserted into equation (2) to compute the Visual PageRank.

Other interesting information emerged from the observations of user browsing behavior. For example, we noticed that users are less likely to click on advertisement images (like banners) than on other image links. On the other hand, users tend to click more on company logos. Even using visual information, our system was not able to perform the sophisticated page analysis needed to distinguish between advertisement image links and company logos. Thus, we decided to discard this potentially useful information.

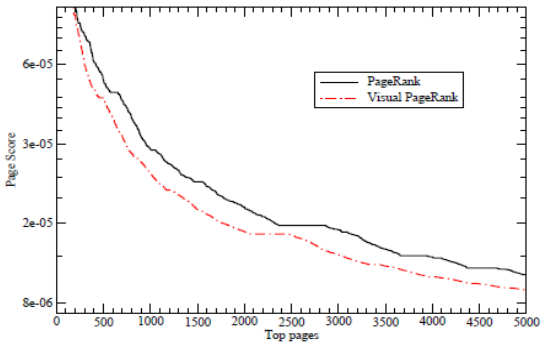


***Figure 2*: The score distribution of the set of top 5000 pages: the visual rank yields a less smooth and more "discriminative" score distribution.**

6. EXPERIMENTAL RESULTS

We performed a set of experiments to compare how the Visual PageRank differs with respect to the standard PageRank. Using the focus crawler described in [2] we downloaded 150.000 documents related to the topic "wine". We then computed the PageRank and the Visual PageRank using the surfer parameters estimated as described in the previous section, on this focused portion of the Web.

We expected the visual information to prevent the rank from flowing through not important links, while increasing the rank flow through the most relevant links in each single page. A first confirmation of this effect is shown in the plot in figure 2, where it is plotted the relevance of the top 5000 documents for the Visual and standard PageRank. The Visual PageRank is less smooth assigning higher values to a smaller subset of pages.

Table 3 shows links of 20 pages with the highest score. Using the Visua] PageRank all top results were relevant for the topic "wine", whereas using standard PageRank some pages that are not authoritative for the considered topic showed up in the first positions, i.e. "www.microsoft.com/ie/logo.asp" and "www.microsoft.com/windows/mediaplayer/download". Even if Visual PageRank is not focused on any specific topic (the content of the page is never taken into account), it is able to not assign a high score to these pages, even if they are linked from a high number of pages. This happens because such links are usually not part of the page core and they are not intended to be immediately seen by the user.

7. CONCLUSION

Visual information contained in a web page is usually discarded. However, this information is very important for a better understanding of the content of web documents. In this paper, the visual information was extracted for any given page and used to predict the behavior of a real user when visiting that page. This prediction has been shown to be useful to generate an improved random surfer model and, as a consequence, a more precise ranking function for search engines.

In spite of the improvement, there is still a big gap between the complex interaction that a real user performs with a page and the simple user model that was employed in our experiments. For example, our experiments showed that real users are less likely to click on an advertisement image links (like banners) than on other image links. On the other hand, users tend to click more on company logos. Even using the visual information, the proposed system is, for the moment, unable to perform the sophisticated page analysis that could allow distinguishing between advertisement image links and company logos. Another possible improvement is to make the rendering module to take into account advanced features like layers and style sheets which are actually ignored, discarding some useful visual information for the most complex page layouts.

We also plan to apply the proposed framework to focused versions of PageRank like that proposed in [5, 8].

Finally, we plan to test the proposed ranking algorithm on a bigger dataset and to measure the ranking accuracy from the feedback of real users.

Please note that the visual approach presented in this paper is very general and it could be used to improve many other applications like page classification and clustering, information extraction, focus crawling and others.

ACKNOWLEDGMENT

We thank Chiara Pintucci from the "Dipartimento di Scienze della Comunicazione" of the University of Siena who performed the user observations that allowed to create the user model used in the paper.

PageRank	Visual PageRank
www.winerelease.com/useragreement.htm	www.winespecialist.com
www.winespecialist.com	www.wineathomeit.com
www.wineathomeit.com	www.connollyswine.co.uk
www.singlesites.com	wineportal.itswine.com
www.closmimi.com	www.connollyswine.co.uk/crt_Cart.asp
www.erobertparker.com/support/...	www.sbwines.com
www.thevirginiacompany.com	forum.wine.co.za/display...
www.connollyswine.co.uk/crt_Cart.asp	www.thevirginiacompany.com
www.connollyswine.co.uk/prodsearch.asp	www.hawleywine.com
www.erobertparker.com/subscriptions/	www.wineanorak.com/taste.htm
www.wineanorak.com/taste.htm	www.ravenswoodlane.com.au
microsoft.com/windows/mediaplayer/...	www.montrosechina.com
www.liquidasset.com	annex.winehouse.com
macromedia.com/shockwave/download/...	www.liquidasset.com
www.adremote.timeinc.net/click.ng...	www.wineanorak.com
www.montrosechina.com/include_html/...	www.clicquot.com
annex.winehouse.com/past.reviews.html	www.cellarexchange.com
www.microsoft.com/ie/logo.asp	www.santabarbara.winecountry.com

***Table 3*: Top 20 score pages for the dataset "wine" when using PageRank and Visual PageRank. Note two irrelevant links from WWW.MICROSOFT.COM website when classical PageRank is used.**

REFERENCES

[1] L. M. Bernard. Criteria for optimal web design (designing for usability). Technical report, Wichita State University, available online at http://psychology.wichita.edu/optimalweb/position.htm, 2001.

[2] M Diligenti, F. Coetzee, S. Lawrence, L. Giles, and M. Gori. Focus crawling by context graphs. In *Proceedings of the International Conference on Very Large Databases*, 11-15 September 2000, Cairo, Egypt, pages 527-534, 2000.

[3] M. Diligenti, M. Gori, and M. Maggini, A unified probabilistic framework for web page scoring systems. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(1):4-16. January 2004.

[4] P. Domingos and M. Richardson. The intelligent surfer: Probabilistic combination of link and content information in pagerank. *Advances in Neural Information Processing Systems,* 14:1441-1448, 2002.

[5] T. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th World Wide Web Conference (WWW11)*, Honolulu, Hawaii, May 2002.

[6] F. James. Representing structured information in audio interfaces: A framework for selecting audio marking techniques to represent document structures. Technical report, Ph.D. thesis. Stanford University, available online at http://www.pcd.stanford.edu/frankie/thesis/, 200l.

[7] M. Kovacevic, M. Diligenti, M. Gori, and V. Milutinovic. Recognition of common areas in web page using visual information: a possible application in a page classification. In *Proceedings of International Conference on Data Mining (ICDM)*, pages 250-258. Maebashi City, Japan, 2002. IEEE Press.

[8] L Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bringing order to the web. Technical report, Computer Science Department, Stanford University, 1998.

[9] E. Seneta. *Non-negative matrices and Markov chains*, Springer Verlag, 1981.

Applied Concept Modeling Techniques for Semantic Data Retrieval

Đorđević, Nemanja; Meštrović, Nenad; Mijović, Đorđe; Dimovski, Boris; and Stancin, Sara

Abstract – The *main topic of this paper is a general overview of currently applicable concept modeling approaches for extracting useful data from large quantities of text that can be accessed nowadays. The paper will begin by discussing the main ideas of concept modeling and the usability it can provide in order to retrieve the actual context from raw texts. Paper also includes current implementation overview as well as a few examples that present the usage of the system, as well as mentioning of issues that yet need to be resolved in the future.*

Index Terms – ***concept modeling, semantic, text mining***

1. Introduction

The Internet and the increasing development of powerful computer systems both for personal and professional usage led to unimagined increase in the amount of available data in various forms. While database systems evolved the capability to provide flexible, fast, and scalable data management, the problems with large texts as data storage units did not diminish. On contrary, with emerging of fast Internet search systems such as Google®, that provide mechanisms to search a vast amount of data syntactically, the need to semantically structure the text has even more decreased. However, the need for semantic text mining did not decrease and users lose much more time doing semantic searches via text-based search tools, which can sometimes be very inefficient. One solution is to the problem is semantic text mining. Semantic text mining is a technique that allows us to create meaningful relationships between abstractions within the text. The created relationships are in fact representations of brand new abstractions. In this paper, we will often refer to these abstractions as concepts, although concepts do not have a limitation on a number of abstractions that they bind. A concept is virtually any meaningful abstraction, which can exist without dependencies towards other abstractions. The idea of the project is to create and extract as many concepts from texts as possible, as long as they deserve to be called a concept (i.e. to have a meaning). In the remainder of the paper we will introduce some observations that make the concept creation process possible and usable.

2. Text characteristics overview and its influence on concept modeling

Generally, texts come in various forms. Besides conforming to language standards like grammar and vocabulary there is nothing else that can be used to structure a text in some generic manner. However there is one characteristic that is common to almost all known world languages, and that is the following: abstractions that form a concept are related not only by meaning, but by distance of words within the text. Concepts are therefore made of abstractions, or better say words, that are adjacent. The idea of concept modeling is to establish connections not only between abstractions to form concepts, but to form connections between concepts as well. Therefore the idea is not only to search for words that are adjacent to the observed

Manuscript received May, 27, 2009.
Djordjevic, Mestrovic, Mijovic and Dimovski are with the School of Electrical Engineering, University of Belgrade, Serbia.
Sara Stancin is with the University of Ljubljana, Slovenia.

word, but also words that appear in its wider neighborhood. Multiple appearances of a candidate concept increase its chance of being promoted to actual concepts. Although the idea is language-independent, there are some issues that make the problem actually bound to a specific language, or better say, to the structure of a that language. There is always a certain probability that we can extract abstractions that do not have a meaning and therefore do not form a concept. In order to increase our chances of retrieving useful content, the text must undergo a filtration stage. The filtration stage is necessary in order to remove language elements we certainly know that will not help us form true concepts. Most of these elements are binders, propositions, pronouns, separators, numbers etc. Besides increasing the chance of retrieving concepts, the side effect of this stage is also noticed as a performance boost to the complete search process, because fewer words need to undergo further processing. The next chapter will introduce some techniques that are used in the implementation. Note that in this stage of research only double-worded concepts are established.

3. Exploiting the idea

a. Implementation details

In order to achieve the desired effects with the filtering stage mentioned in the text above, we need to provide the engine a list of words that need to be filtered and excluded from the search from that point onwards. This is the part of the solution that is language-dependent, so if there is need to make the engine run for languages other than English, one has to assemble a list of the unwanted words (commonly called stop-words), as mentioned above. This particular implementation is based on English texts only, in order to simplify the research process by using only the standard ASCII character palette. The engine was given a list of around 500 words that need to be excluded. Be aware that the filtering stage as the first stage of the process affects all subsequent stages, and thus have a considerable effect on final results. Failing to properly deduce the stop-words can severely downgrade the quality of the output.

The output of the filtering stage is considered as the actual input to the processing engine. In order to help fulfilling the processing task, a specially designed database is used to store intermediate data. The main elements of this data consist of words that are related by distance, which is their main relationship parameter, and number of appearances of the concept within the text. Also there is a record of the frequency of the concept along various documents. The frequency counting is essential for concept relevance evaluation process. Relevance evaluation has critical impact on the results, and by using appearance counting in two levels (counting against the text alone, and counting against the number of appearances in different documents) the overall quality of the output is dramatically increased. By setting the low-limit to the number of appearances a potential concept needs to breach we can fine-tune the results quality vs. quantity characteristic as we shall see in the example outputs later.

b. Tuning the test

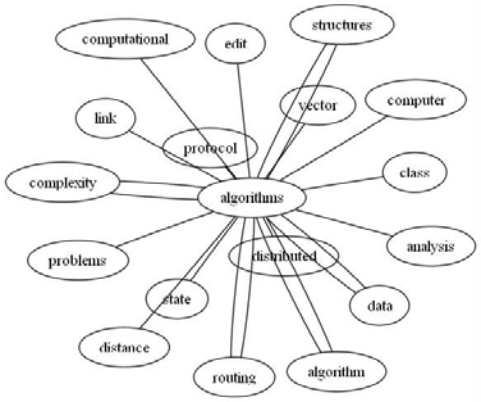
During the white box testing phase, we tested vast amount of documents with different concept distance and appearance limits. It was attempted to establish concepts between words that are up to 10 words away. As expect, this turned out to be much more than needed, as the median distance for all real concepts was about 5.4, though it is important to mention that many concepts were found on great distances, even on 10. The reason for this is the fact that by removing the stop-words, we also delimited sentences, so words from different sentences could have been connected. However, these concepts always appeared much more frequent on lower distances, so the exclusion of these candidates did not degrade the results. Naturally, the same frequency breaching low-limit could not be used for documents that varied greatly in total number of words in text. Therefore it was attempted to establish a connection between a number of words within the text and the frequency low-limit that give optimal results. The targeted optimal result limit was 80%, meaning that only 1 out of 5 concepts

candidates that are promoted to real concepts are allowed to be false. The discovery of function between the number of words and the desired breaching limit turned out to be much more difficult task then expected, because it varied greatly along different types of texts. The testing itself used Wikipedia® as primary source. There are two reasons for this decision. One reason is that these articles were compiled by a number of people, making it much less standardized because no single form is applied to all of the available documents. The second reason is that these texts are suitable for concept modeling as they generally include many concepts and are thus suitable for testing as more useful data can be collected in less time. The low-limit value we managed to agree on was set to 0.5% of total word count as no other function produced better results to justify its use for all the tested documents. This means that for example, the text that consists of 3000 words will have a breaching limit of 15. Someone might say that the limit is too high, but the number of false concepts increases in similar manner to the number of real concepts, so any lower value would threaten the output quality. The linear function turned out to be a simple yet effective solution to the problem.

4. Analyzing the results

The output of the processing is parameterized by minimum number of documents a candidate concept must be found in. This particular test was undergone on Wikipedia® articles that are related to computer science in general. Also, articles related via hyperlinks were processed too, to give a total of 45 documents. Due to the time complexity of the entire processing we were unable to process a larger amount of documents with currently implemented methods and available hardware, but nevertheless, the concept table contained 3118 records which proved to be relevant sample after all. First 3 outputs are produced using “algorithms” keyword while the 4th was an output to keyword “digital”. Results are shown in a form of a graph, connecting related words into concepts.

Test 1

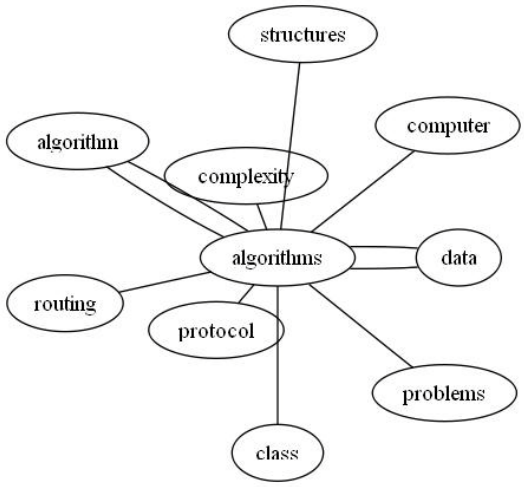


- keyword - “algorithms”
- document low-limit - 1

As you can see, there are some false concepts that were outputted, meaning that the breaching limit is not high enough.

Test 2

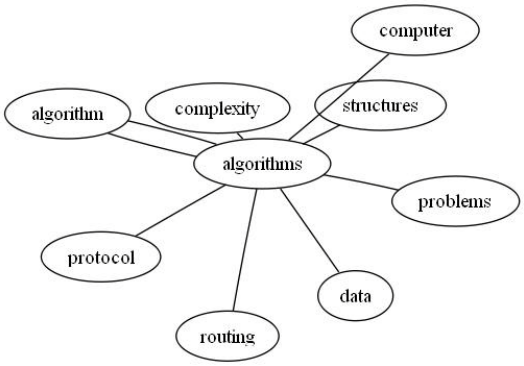
.



- keyword – “algorithms”
- document low-limit – 2

Now we can see an improvement in favor of quality rather than quantity.

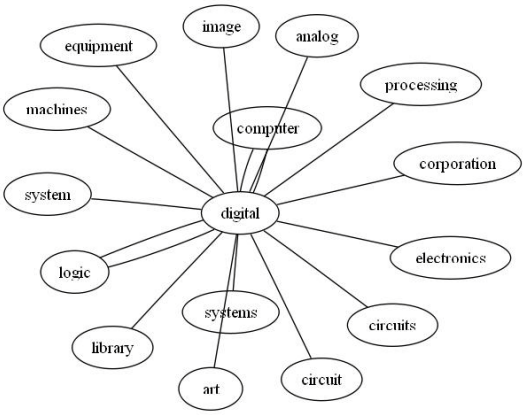
Test 3



- keyword – “algorithms”
- document low-limit – 3

In this case only a small improvement is noticed with the disappearance of the “class” abstraction. As you can see, there is no linear improvement of results with increasing of document low-limit boundary.

Test 4



- keyword – “digital”
- document low-limit – 1

In the last example, we received outstanding results beyond our own expectation, even with minimum document low-limit of 1. Almost all of the connections make real concepts. So, the general conclusion is that the document low-limit parameter has yet more to be examined, and the way it effects the results.

5. Future improvements

The current implementation suffers from many limitations. Many of them are caused by limited hardware resources due to much extended processing times. Although there is still some room for optimizations, it is undoubtedly that the technique itself needs powerful hardware in order to provide accurate results and to forms large concept networks, which would increase the relevance of searches in general. The next generation concept modeling engine will benefit from standards like HTML and XML that provide hyperlinks and custom structuring that can ease relevance determination, providing even better results. Nevertheless, what is essential to the technique is that large quantities of texts need to be processed, and that the quality of the results is maximized by creating a proper list of stop-words.

Although still far away from real-world exploitation, the concept modeling text mining technique will surely be a tool the fast text search engines will find very useful.

Acknowledgement

Authors are thankful to Prof. Tomazic of U. of Ljubljana, for his help.

References

[1] OMEROVIC, Sanida, TOMAZIC, Saso, MILUTINOVIC, Veljko. Concept modelling. V: AURER, Boris (ur.), BACA, Miroslav (ur.), RABUZIN, Kornelije (ur.). 19th Central European Conference on Information and Intelligent Systems, September 24-26, 2008, Varazdin, Croatia. Conference proceedings. Varazdin: University of Zagreb, Faculty of Organisation and Informatics, 2008, str. 33.

[2] OMEROVIC, Sanida, JAKUS, Grega, FILIMONOVA, Tatjana, TOMAZIC, Saso. Zapis vecjezicnih besedil v e-sperantu. Elektroteh. vestn., 2007, letn. 74, st. 4, str. 151-157, ilustr. [COBISS.SI-ID 6001492]

[3] TOMAZIC, Saso. Multilingual web with E-speranto. IPSI BGD Trans. Internet Res.. [Print ed.], Jul. 2007, vol. 3, no. 2, str. 13-15.

Hot Topics in Computer Science and Engineering

Tomažič, Sašo and Milutinović, Veljko

INTRODUCTION

This paper is a result of recent discussions at MIT (Massachusetts Institute of Technology) during which selected professors and/or their assistants presented their views (pearls of wisdom), in front of a group of scientists from all over the world interested in multi, inter, and trans-disciplinary research. Topics in this paper appear in the order in which they were presented.

The paper consists of ten sections on ten important computer science and engineering (CSE) topics. We do not know what the final research results will be; we only believe that we have learned which avenues might lead to results, and we present them, in expectation of the response from the readership.

Each section of this paper includes the following elements: (a) Brief topic description and its social relevance; (b) Pointer to a state-of-the-art research at MIT; (c) Technical challenge on the way to a more effective future solution; (d) Important note or something to keep in mind while studying the subject.

The major purpose of this paper is education of graduate students in the topics of mission and vision of further advances in computing and communications.

I. GLOBALIZED EDUCATION

Globalized world can be achieved only if global education is performed in all relevant languages of the world, as underlined in the research of Tom Greene. A tool that enables automatic translation of courseware among different languages can be treated as an important accelerator of the globalization process. Each individual on earth can start thinking creatively only if properly educated, with a stress on creative thinking.

An important state-of-the-art solution in the field is the Open Courseware of MIT, with about 1400 different free courses, as well as the related translation efforts (no intellectual property issues involved), as an important step in the process.

The mayor technological challenges for future research are translation and motivation efforts based on NLU (natural language understanding) and system integration and optimization efforts based on PSN (psycho social neurology).

Information about everything is doubling every two years and it now takes 10 years to develop experts. To be truly expert, they must also learn how to learn to find the answers for what they do not know by learning and using the best, most current information.

II. QUICK REUSE OF EXISTING DESIGNS

Effective tools are needed for fast design of Application-Specific Integrated Circuits (ASICs). The variety of chips must increase, so we can have a variety of applications at the ASIC power consumption of typically 1W, rather than the PC power consumption of typically 100W. Such a power dissipation reduction enables the penetration of sophisticated applications into the current cell-phone environment or the future wearable computing environments.

The MIT Term Rewriting Systems (TRS) based synthesis technology, which is in the foundation of BlueSpec Electronic Design Automation (EDA) toolset, enables quick and reliable design of new chips as well as quick reuse of existing designs.

The mayor challenge is building a library of micro-architectural Intellectual Property (IP) blocks and how IP blocks should be written to

Manuscript received May, 5th, 2009.

Veljko Milutinovic is with the School of Electrical Engineering, University of Belgrade, Serbia.

Saso Tomazic is with the Faculty of Electrical Engineering, University of Ljubljana, Slovenia.

be easily modifiable and reusable, as in the research efforts leaded by Arvind.

The major issues are power, size, programmability, and time to market in less than 6 months. After 12 months there is no profit and after 18 months some new technology comes that brings new applications.

III. Hardware Helping Software

Single processor computers are reaching their limit. Computers with multiple processors will rule in the future. However, programming for multiple processors is still a disaster area due to unwanted asynchronous hazards. An appropriate hardware architecture is needed that makes the software less likely to be incorrect (there is no software failure – software can only be incorrect).

State-of-the-art is that industry knows how to put very many CPUs on a single chip, but nobody knows how to program them. Different languages for programming multiple processors, as for example, Sisal from Lawrence Livermore National Laboratory, and different research efforts, as for example, Monsoon data flow multiprocessor lead by Arvind, address this problem.

One of the avenues leading to possible solution is to make computer architecture that employ simultaneous multithread processors, employ a global shared name space, and support modular program construction through functional programming, as in the research of Fresh Breeze multi-core architecture leaded by Jack Dennis.

It makes no sense to design computers without taking into account the principles of software methodology and program language design.

IV. Programming Sensor Networks

Appropriate languages are needed for effective programming of systems with millions of processors, in general, and in particular in sensor networks. In many sensor network applications, the network is deployed to approximate a physical space. For example, millions of sensors can be thrown from a plane to a field to measure humidity. They must communicate with each other and with the man processor. Although many of them can be damaged, the network as a whole must still perform its job.

A state-of-the-art solution to the problem is Proto, a high-level language for programming sensor networks developed by Bachrach and his group. The spatial nature of sensor network is abstracted as amorphous medium. This separates sensor network problems into two largely independent domains. Above the abstraction barrier we are concerned with description of applications, while below the barrier we are concerned with fast, efficient, and robust communication between neighboring devices. Existing applications, such as target tracking and threat avoidance, can be expressed in only a few lines of Proto code.

The main challenge is programming of self-organizing systems. The study of self-organizing systems has now reached the tool-building phase, in which a new discipline of self-managing systems engineering can begin to emerge. The next step is to refine the principles of self-organization into a system of parts suitable for engineering – as the principles of electromagnetism are captured for electronic engineering in components like capacitors, transistors and resistors.

The ability to control emergent phenomena depends on decomposing them into aspects susceptible to independent engineering. For spatial self-managing systems, the amorphous-medium abstraction lets you separate the system specification from its implementation.

V. Placing Intelligence into a System of Devices

Intelligence has to be placed into a system of devices; not into a single device. At hock networks formed from consumer electronic devices, as cell phones, PDAs, notebooks, cameras, etc., which can perform certain tasks which could not be accomplished with a single device.

An example of such ad hoc network developed in MIT Media Laboratory lead by Bove is a network formed by personal devices of people present at a meeting, which can form a microphone array. The devices use an active audio self-localization algorithm which is effective in distributed sensor networks even if only coarse temporal synchronization can be established.

Another distributed algorithm is then used to recognize the active speaker and electronically "turn" the microphone to his direction.

Often the amount of local data produced by these devices, and their sheer number, makes centralized data processing infeasible. Many distributed algorithms exist that work locally to produce results from a collection of nodes, but as this number grows the algorithm's performance is quickly crippled by the resulting exponential increase in communication overhead. One possible solution of this problem is local organization of nodes by a protocol that provides a method for nodes to create a bottom-up organization based purely on local conditions. Another is in programming model based on mobile code fragments which self assemble into larger structures.

By connecting personal resources with those of the people nearby, tasks can be achieved, through distributed placement and statistical improvement that a single device could not do alone.

VI. Placing Intelligence into Robots

In robotics a point was reached where it will be possible to build robots that will act as partners and not only industrial assemblers or children toys. They will act as teachers, domestic assistants, healthcare assistants, and so on. A great emphasis will be given to Human-Robot Interface (HRI).

State-of-the art robot is Leonardo. He has 61 degrees of freedom – 32 of those are in the face alone. As a result, Leonardo is capable of near-human facial expression.

Robotic Life Group at MIT leaded by Cynthia Breazeal takes inspiration and guidance from the science of animal and human behavior to build cooperative robots that can work and learn in partnership with people. The main challenge is to teach a robot as you teach a child, through guidance with rewards and punishments.

The research of human-like robots is multidisciplinary in nature. It can not only help to build robots with human-like behavior but can also give scientific insight into the mechanisms that underlie human competence and deeper understanding of human abilities.

VII. Placing Intelligence into Homes

Computational sensing and Artificial Intelligence (AI) techniques – integrated with architectural principles and effective interface design – can be used to create new types of living environments which would greatly improve Quality of Life (QoL).

The PlaceLab – an apartment-scale shared research facility of MIT Architecture research group leaded by Stephen S. Intille – is a state-of-the-art tool for exploring how new technologies, materials, and design concepts can make possible dynamic, evolving places that respond to the complexities of life. Hundreds of sensing components are installed in nearly every part of the department. The facility is managed as a multi-disciplinary shared scientific tool in the tradition of other scientific facilities developed to study unique environments.

A special focus is now given to new algorithms that use ubiquitous sensing to recognize activities of daily living in order to determine when and how to provide preventive healthcare education in the home. Main challenges are: How can ubiquitous computing technologies be created and evaluated that work for extended periods of time? How can persuasive interfaces designed to motivate behavior change in the home be studied in context?

Measuring behavior can change behavior when a person is aware that he/she is monitored. She/he starts to behave differently, to avoid socially unaccepted behaviors, etc. Like in quantum mechanics, certain things simply cannot be measured.

VIII. Placing Intelligence Everywhere

Once the home infrastructure has become intelligent, the next step is to inject intelligence into all kinds of everything at homes beyond homes. People have developed sophisticated skills for sensing and manipulating our physical environments. However, most of these skills are not employed by traditional Human Computer Interface (HCI).

State-of-the-art examples from the MIT Media Lab Tangible Media Group include lukastro suits (textile, iron, cotton, nylon) that detect touches (for applications like home and workplace abuse detection), an electrically

charged coat that could prevent women from being attacked, domestic applications with funny interfaces (e.g., Blendie, an interactive, sensitive, intelligent, voice controlled blender with a mind of its own), monitoring of the environment and changing of personal trajectories for risk reduction (e.g. changing of song tunes for joggers in heavy traffic), or intelligent Lego cubes (to enhance creativity in children).

The Computing Culture group lead by Chris Csikszentmihályi creates unique media technologies for cultural applications. Projects will result in specific works of art, but will also help further an understanding of the relationships between new media and cultural production. Some of the strategies they utilize include interventions in contemporary consumer electronics, creating special events for public situations, and applying technical research and development to cultural agendas that wouldn't normally receive them.

Future efforts must involve artists to help reconfigure technology, to address the full range of human experience. Of special importance are the issues related to creativity enhancement in children.

IX. Web of Trusted Resources and Web for Everyone

By the words of the director of W3C and inventor of the World Wide Web Sir Tim Berners-Lee: “The social value of the Web is that it enables human communication, commerce, and opportunities to share knowledge. One of W3C's primary goals is to make these benefits available to all people, whatever their hardware, software, network infrastructure, native language, culture, geographical location, or physical or mental ability.” However, for Web to be a useful, people must be able to trust the resources. While technology cannot guarantee trust, it should enable secure transactions with trusted parties. The W3C has already developed its agenda for future developments.

World Wide Web has already reached 16.6% global user penetration explained Steve Bratt, the Chief Executive Officer of W3C. The Web has transformed the way we communicate with each other. People now "meet on the Web" and carry out commercial and personal relationships, in some cases without ever meeting in person.

One of W3C's goals is to design technology that will work independent of a particular hardware platform, including mobile telephones, kiosks in airports, kitchen appliances, and automobiles. Access from these devices should be as simple, easy and convenient as Web access from a home computer.

While foundation of today's Web is in HTML, CSS, XML, and Architecture, foundation for tomorrow’s Web is in Web 2.0, Web services, Semantic Web, Mobile Web, ubiquity, universality, internationalization, security and privacy.

X. Getting an Extremely fast Computing Infrastructure

The ultimate goal in concept modeling (which is the key issue for a number of important applications of the future) is how to get closer to the brain model. This implies an extremely strong computing infrastructure. The driving forces towards a much stronger computing power are not only the needs of ubiquitous computing and web of trusted resources, but also the needs of fundamental research in particles physics.

State-of-the-art needs one billion collisions in an event per second, and about 200 to 300 particles per event. Today, this is achievable only off-line.

This to be achievable on-line believes Nobel Laureate Jerome I. Friedman. That implies zillions of CPUs talking to each other. Once this computing/communications goal is achieved, we are very close to the brain model. Computer communications is like cheering and drinking. If two persons like to cheer and drink, they can do so many times in one evening, which means that they can drink many glasses of wine during that evening. Zillions persons, if they all like to cheer with each other, will not drink a single glass of wine in their entire lifetime.

Note that often technological advances in one specific field may trigger application advances in another field, less or more remote. On the other hand, scientists start making errors when conditions change – subconsciously each person likes to protect the time he/she spent on learning something. The major ad-

vances are usually made by researchers which novel to the field, who do not know that it cannot be done. Many Nobel Laureates were told, by their senior colleagues, that what they are doing is a waste of time.

Conclusion

From the above presentations we conclude that lot of effort is given to improve quality of life, proper education accessible to every one in his own language, being one of the most important prerequisite to obtain this goal. Only after this goal is achieved, different advances in technology can make life easier and more pleasant for everyone on this Earth.

From technological point of view, in computer science, standalone single computer systems became to complex and to power consumptive. The only alternative are systems with many, possible millions, of units than can communicate among each other and perform complex tasks which would be unthinkable of stand alone computers. The idea resembles the way the nature “thinks”, using many very simple and very similar units (e.g., cells), to “produce” very complex systems (e.g., human beings).

We also noticed that the research in computer science is becoming more and more multi, inter, and trans disciplinary, as scientists from almost all fields must be included in development of modern, computer added systems at one hand and the researches in all fields need some form of computers as a tools in their research at the other hand. Thus, people with creative mind and wide horizons which are capable of horizontal thinking will be the pillows of further progress in the field.

ACKNOWLEDGEMENT

This paper is based on presentations given by professors and researchers at MIT:

- Thomas J. Greene, MIT Computer Science and Artificial Intelligence Laboratory, outreach officer
- Arvind, MIT Department of Electrical Engineering and Computer Science, Johnson Professor
- Jack B. Dennis, MIT Department of Electrical Engineering and Computer Science, Professor Emeritus
- Jonathan R. Bachrach, MIT Artificial Intelligence Laboratory, Research Scientist.
- V. Michael Bove, Jr., MIT Media Lab, Head of Consumer Electronics Laboratory, Principal Research Scientist
- Cynthia Breazeal, MIT Media Lab, Head of Robotic Life group, Associate Professor of Media Arts and Sciences
- Cory D. Kidd, MIT Media Lab, Robotic Life group, PhD Student.
- Stephen S. Intille, MIT Department of Architecture, Research Scientist and Technology Director, House_n Consortium
- Chris Csikszentmihályi, MIT Media Lab, Head of Computing Culture group, Associate Professor of Media Arts and Sciences.
- Kelly Dobson, MIT Media Lab, Computing Culture group, PhD Student.
- Adam Whiton, MIT Media Lab, Tangible Media Group, Industrial Designer
- Steve Bratt, World Wide Web Consortium, Chief Executive Officer.
- Jerome I. Friedman, MIT Laboratory for Nuclear Science, Institute Professor and Professor of Physics, 1990 Nobel Laureate.

The authors of this paper would like to thank them and all the others, for taking their time for these very educational presentations.

Reviewers by Countries

Argentina
Olsina, Luis; National University of La Pama
Ovando, Gabriela P.; Universidad Nacional de Rosario
Rossi, Gustavo; Universidad Nacional de La Plata
Australia
Abramov, Vyacheslav; Monash University
Begg, Rezaul; Victoria University
Bem, Derek; University of Western Sydney
Betts, Christopher; Pegacat Computing Pty. Ltd.
Buyya, Rajkumar; The University of Melbourne
Chapman, Judith; Australian University Limited
Chen, Yi-Ping Phoebe; Deakin University
Hammond, Mark; Flinders University
Henman, Paul; University of Queensland
Palmisano, Stephen; University of Wollongong
Ristic, Branko; Science and Technology Organisation
Sajjanhar, Atul; Deakin University
Sidhu, Amandeep; University of Technology, Sydney
Sudweeks, Fay; Murdoch University
Austria
Derntl, Michael; University of Vienna
Hug, Theo; University of Innsbruck
Loidl, Susanne; Johannes Kepler University Linz
Stockinger, Heinz; University of Vienna
Sutter, Matthias; University of Innsbruck
Walko, Zoltan
Brazil
Parracho, Annibal; Universidade Federal Fluminense
Traina, Agma; University of Sao Paulo
Traina, Caetano; University of Sao Paulo
Vicari, Rosa; Federal University of Rio Grande
Belgium
Huang, Ping; European Commission
Canada
Fung, Benjamin; Simon Fraser University
Grayson, Paul; York University
Gray, Bette; Alberta Education
Memmi, Daniel; UQAM
Neti, Sangeeta; University of Victoria
Nickull, Duane; Adobe Systems, Inc.
Ollivier-Gooch, Carl; The University of British Columbia
Paulin, Michele; Concordia University
Plaisent, Michel; University of Quebec
Reid, Keith; Ontario Ministry og Agriculture
Shewchenko, Nicholas; Biokinetics and Associates
Steffan, Gregory; University of Toronto
Vandenberghe, Christian; HEC Montreal
Croatia
Jagnjic, Zeljko; University of Osijek
Czech Republic
Kala, Zdenek; Brno University of Technology
Korab, Vojtech; Brno University of technology
Lhotska, Lenka; Czech Technical University
Cyprus
Kyriacou, Efthyvoulos; University of Cyprus
Denmark
Bang, Joergen; Aarhus University
Edwards, Kasper; Technical University Denmark
Orngreen, Rikke; Copenhagen Business School
Estonia
Kull, Katrin; Tallinn University of Technology
Reintam, Endla; Estonian Agricultural University
Finland
Lahdelma, Risto; University of Turku
Salminen, Pekka; University of Jyvaskyla
France
Bournez, Olivier
Cardey, Sylviane; University of Franche-Comte
Klinger, Evelyne; LTCI – ENST, Paris
Roche, Christophe; University of Savoie
Valette, Robert; LAAS - CNRS
Germany
Accorsi, Rafael; University of Freiburg
Glatzer, Wolfgang; Goethe-University
Gradmann, Stefan; Universitat Hamburg
Groll, Andre; University of Siegen
Klamma, Ralf; RWTH Aachen University
Wurtz, Rolf P.; Ruhr-Universitat Bochum
Greece
Katzourakis, Nikolaos; Technical University of Athens
Bouras, Christos J.; University of Patras and RACTI
Hungary
Nagy, Zoltan; Miklos Zrinyi National Defense University
India
Pareek, Deepak; Technology4Development
Scaria, Vinod; Institute of Integrative Biology
Shah, Mugdha; Mansukhlal Svayam
Ireland
Eisenberg, Jacob; University College Dublin
Israel
Feintuch, Uri; Hadassah-Hebrew University
Italy
Badia, Leonardo; IMT Institute for Advanced Studies
Berrittella, Maria; University of Palermo
Carpaneto, Enrico; Politecnico di Torino
Japan
Hattori, Yasunao; Shimane University
Livingston, Paisley; Linghan University
Srinivas, Hari; Global Development Research Center
Obayashi, Shigeru; Institute of Fluid Science, Tohoku University
Mexico
Morado, Raymundo; University of Mexico
Netherlands
Mills, Melinda C.; University of Groningen
Pires, Luís Ferreira; University of Twente
New Zealand
Anderson, Tim; Van Der Veer Institute

Philippines
Castolo, Carmencita; Polytechnic University Philippines
Poland
Kopytowski, Jerzy; Industrial Chemistry Research Institute
Portugal
Cardoso, Jorge; University of Madeira
Natividade, Eduardo; Polytechnic Institute of Coimbra
Oliveira, Eugenio; University of Porto
Republic of Korea
Ahn, Sung-Hoon; Seoul National University
Romania
Moga, Liliana; "Dunarea de Jos" University
Serbia
Mitrovic, Slobodan; Otorhinolaryngology Clinic
Stanojevic, Mladen; The Mihailo Pupin Institute
Ugrinovic, Ivan; Fadata, d.o.o.
Singapore
Tan, Fock-Lai; Nanyang Technological University
Slovenia
Kocijan, Jus; Jozef Stefan Institute and University of Nova Gorica
South Korea
Kwon, Wook Hyun; Seoul National University
Spain
Barrera, Juan Pablo Soto; University of Castilla
Gonzalez, Evelio J.; University of La Laguna
Perez, Juan Mendez; Universidad de La Laguna
Royuela, Vicente; Universidad de Barcelona
Vizcaino, Aurora; University of Castilla-La Mancha
Vilarrasa, Clelia Colombo; Open University of Catalonia
Sweden
Johansson, Mats; Royal Institute of Technology
Switzerland
Niinimaki, Marko; Helsinki Institute of Physics
Pletka, Roman; AdNovum Informatik AG
Rizzotti, Sven; University of Basel
Specht, Matthias; University of Zurich
Taiwan
Lin, Hsiung Cheng; Chienkuo Technology University
Shyu, Yuh-Huei; Tamkang University
Sue, Chuan-Ching; National Cheng Kung University
Ukraine
Vlasenko, Polina; EERC-Kyiv

United Kingdom
Ariwa, Ezendu; London Metropolitan University
Biggam, John; Glasgow Caledonian University
Coleman, Shirley; University of Newcastle
Conole, Grainne; University of Southampton
Dorfler, Viktor; Strathclyde University
Engelmann, Dirk; University of London
Eze, Emmanuel; University of Hull
Forrester, John; Stockholm Environment Institute
Jensen, Jens; STFC Rutherford Appleton Laboratory
Kolovos, Dimitrios S.; The University of York
McBurney, Peter; University of Liverpool
Vetta, Atam; Oxford Brookes University
Westland, Stephen; University of Leeds
WHYTE, William Stewart; University of Leeds
Xie, Changwen; Wicks and Wilson Limited
USA
Bach, Eric; University of Wisconsin
Bazarian, Jeffrey J.; University of Rochester School
Bolzendahl, Catherine; University of California
Bussler, Christoph; Cisco Systems, Inc.
Charpentier, Michel; University of New Hampshire
Chester, Daniel; Computer and Information Sciences
Chong, Stephen; Cornell University
Collison, George; The Concord Consortium
DeWeaver, Eric; University of Wisconsin - Madison
Ellard, Daniel; Network Appliance, Inc
Gaede, Steve; Lone Eagle Systems Inc.
Gans, Eric; University of California
Gill, Sam; San Francisco State University
Gustafson, John L..; ClearSpeed Technology
Hunter, Lynette; University of California Davis
Iceland, John; University of Maryland
Kaplan, Samantha W.; University of Wisconsin
Langou, Julien; The University of Tennessee
Liu, Yuliang; Southern Illinois University Edwardsville
Lok, Benjamin; University of Florida
Minh, Chi Cao; Stanford University
Morrissey, Robert; The University of Chicago
Mui, Lik; Google, Inc
Rizzo, Albert ; University of Southern California
Rosenberg, Jonathan M. ; University of Maryland
Shaffer, Cliff ; Virginia Tech
Sherman, Elaine; Hofstra University
Snyder, David F.; Texas State University
Song, Zhe; University of Iowa
Wei, Chen; Intelligent Automation, Inc.
Yu, Zhiyi; University of California
Venezuela
Candal, Maria Virginia; Universidad Simon Bolívar

IPSI Team
Advisors for IPSI Developments and Research:
Zoran Babovic, Darko Jovic, Aleksandar Crnjin, Marko Stankovic, Marko Novakovic

Authors of papers are responsible for the contents and layout of their papers.

Welcome to IPSI BgD Conferences and Journals!

http://www.internetconferences.net

http://www.internetjournals.net

VIPSI-2010 AMALFI
Hotel Santa Caterina
Amalfi, Italy
March 4 to March 7, 2010

VIPSI-2010 TIVAT
Tivat, Montenegro
June 4 to June 7, 2010

VIPSI-2010 TIVAT
Tivat, Montenegro
October 1 to October 4, 2010

VIPSI-2010 MILOCER
Milocer, Montenegro
December 30 to December 31, 2010

CIP – Katalogizacija u publikaciji
Narodna biblioteka Srbije, Beograd
ISSN 1820 – 4503 =
The IPSI BGD Transactions on Internet Research
COBISS.SR - ID 119128844