

Centre for Next Generation Localisation (<http://www.cngli.ie/>) in conjunction with IBM Ireland, are planning to run a tutorial Monday January 19 – Wednesday January 21 2009 in Dublin City University on:

Graph-based Mining of Digital Content

Tutorial summary

Proliferation of Web 2.0 and Web 3.0 creates massive computer mediated networks. It is expected that by the year 2010, most of the Web information will be created automatically as logs files of Web services (such as Facebook) and by “The Internet of Things“. Graphs serve as suitable models for such multidimensional networks where human generated content is part of larger networks. This tutorial will present novel methods of graph-based mining of large volumes of information.

Tutorial content

The tutorial will cover the following topics

Spreading Activation Framework for Graph Mining

Spreading activation (also called as spread of activation) is a method for searching associative networks, neural networks or semantic networks. The method is based on the idea of quickly spreading an associative relevancy measure over the network. Our goal is to give an expanded introduction to the method. We will demonstrate and describe in sufficient detail that this method can be applied to very diverse problems and applications. We will present this method as a very general class of algorithms on large (or very large) multidimensional networks which will serve a mathematical model. We will also outline applications of this framework for ontology-based text processing, for application in collaborative tagging systems, and for mining of massive dynamic social networks.

Fuzzification of Graph Clustering Algorithms and Machine-based Clustering with Learning for Large Document Bases

Among various algorithms that have been proposed so far for clustering digital contents, the most popular are those that automatically reveal the number of clusters and assign each target document to exactly one cluster. However, in many real situations, an exact boundary between different clusters of digital contents does not exist. We introduce a fuzzy version of the MajorClust algorithm. The proposed clustering method assigns documents to more than one category by taking into account a membership function for both, edges and nodes of the corresponding underlying graph.

Machine-based Localization of Hidden Knowledge in the Internet

Design of fast real-time algorithm for the optimal search for the hidden information in the Internet documents is a hot area of research in mining digital contents. We consider a new family of algorithms based on a priori expert information: the probability of finding a needed information in each specific Internet document, probability of an unsuccessful search, cost and time of search trials. The concept of the integral utility, or fitness, of the documents is precisely defined. The necessary and sufficient optimality conditions for the local search to be globally optimal are found which claim that only linear and exponential utility functions has this property.

Social Semantic Desktop Applications of IBM library Galaxy

IBM Library Galaxy (<http://www.alphaworks.ibm.com/tech/galaxy>) is an integral part of the social semantic desktop created by the EU project Nepomuk (<http://nepomuk.semanticdesktop.org/>). In the Nepomuk Networked Environment for Personal Ontology-based Management of Unified Knowledge, Galaxy is used for semantic analysis of text and as a generic recommender for networked data. This tutorial will provide a starter's guide how to install the Nepomuk, how to use Galaxy for text analytics, for link analysis, and for mining of socio-semantic networks.

Presenters

The tutorial will be a collaborative effort between a number of researchers from several institutions:

- IBM Ireland Centre for Advanced Studies and IBM LanguageWare group
- Holon Institute of Technology and Bar-Ilan University, Israel.
- EU project Nepomuk partners, including DERI Galway, DFKI, and KTH
- Telecommunications Software & Systems Group of Waterford Institute of Technology (Ireland)

The main contact for the tutorial will be Dr. Alexander Trousov of IBM (e-mail: atrousso@ie.ibm.com, homepage: <http://atroussov.com/>). CNGL Education and Outreach Manager is Cara Green (email: cgreene@computing.dcu.ie).

Prof. Eugene Levner

Eugene Levner is Professor of Computer Science at Holon Institute of Technology and Bar-Ilan University, Israel. His main scientific interests are design of computer algorithms, optimization theory, and clustering and classification of digital content. He is author/co-author of seven books and more than 100 articles in refereed journals. His Citation Index is 410, and h-index is 15. He is the full member of the International Academy of Information Sciences, a member of editorial boards of four international journals.

Dr. Alexander Trousov

Dr. Alexander Trousov is chief scientist in IBM Ireland Centre for Advanced Studies (CAS) and chief scientist of IBM LanguageWare group. He has published more than 30 peer reviewed journal and conference papers and has 5 patents. In 2000 he joined IBM as the Architect of IBM Dictionary and Linguistic tools group, known now as IBM LanguageWare group. As CAS Chief Scientist, Dr. Alexander Trousov leads IBM Ireland's participation in the 3 year integrated 6th framework EU project NEPOMUK, and is one of the creators of IBM Galaxy library.

Dr. John Judge

Dr. John Judge is a researcher in the IBM Dublin Software Lab working as part of the LanguageWare team. He previously worked in the National Centre for Language Technology in Dublin City University, where he received his Ph.D. He joined IBM in 2006 as part of the LanguageWare Team. His current duties include research and development work for IBM's participation in a 3 year integrated EU project NEPOMUK, where the IBM team is developing tools for semantic analysis of text.

Contributors

Mikhail Sogrin (IBM, Ireland)
Dr. Dmitri Borvich (WIT, Ireland)
Dr. Cristian Bogdan (KTH, Ireland)