Carlos Castillo
and Yiyu Yao

EvalWare: Granular Computing for Web Applications

Please send suggestions for Web resources of interest to our readers, proposals for new topics, and general feedback, by e-mail to Alen Docef ("Best of the Web" Associate Editor) at adocef@vcu.edu.

In this issue, "Best of the Web" focuses on granular computing (GC) and applications. GC makes use of multiple levels of granularity in problem solving and draws ideas from areas such as fuzzy computation, evolutionary computation, and machine learning. Emerging applications of GC include data mining, data clustering, classification and aggregation; modeling of cognitive processes; and ontology learning. When applied to signal processing on the World Wide Web (WWW), GC explores the structures, semantics, and knowledge of the Web by a systematic investigation at multiple levels of abstraction. In what follows we focus on resources that are relevant for GC and its potential application to Web data mining (search, retrieval), navigation (crawling), and spam filtering.

For the resources presented here, the attributes in square brackets describe the types of information. Unless otherwise noted, the resources are free. This resource list is also available by convenient point and click on the *IEEE Signal Processing Magazine* Web site at http://apollo.ee.columbia.edu/spm/?i=external/bow.

GC RESOURCES

TUTORIALS

http://en.wikipedia.org/wiki/Granular_computing
http://en.wikipedia.org/wiki/Fuzzy_logic
http://en.wikipedia.org/wiki/Rough_set
http://wi-consortium.org/
[tutorials]

In the title of his talk "Granular Computing: Computing with Information That Is Imprecise, Uncertain, Incomplete, or Partially True" at the 2007 *IEEE International Conference on GC*, Lotfi A. Zadeh reminded everyone the basic definition of GC.

Further details and introductory material on GC are available, some of which (such as that on Wikipedia) explain basic concepts of GC in terms of rough set terminologies. This is not surprising since most of the current research in GC is dominated by set-theoretic models such as rough sets (proposed by Pawlak in the 1970s and published in a book in 1991) and fuzzy sets (proposed by Lotfi A. Zadeh in 1965).

Applications of these and GC in the Web context are available on the site of the Web Intelligence Consortium (WIC), which contains white papers, tutorials and a list of WIC centers around the world.

PORTALS

http://www.granular-computing.org/
http://roughsets.home.pl/www/
http://web.abo.fi/~rfuller/fuzs.html
http://www.fmi.uni-stuttgart.de/fk/evolalg/
[one-stop information portals]

A recently created (July 2007) Web site devoted to GC has become popular in its short existence as an information center that contains news, lists of researchers and publications in GC.

Web search. Cartoon by Tayfun Akgul (tayfun.akgul@ieee.org)

Digital Object Identifier 10.1109/MSP.2007.914239

1053-5888/08/$25.00©2008IEEE

The International Rough Set Society (IRSS) Web site makes available many useful materials, including news, lists of researchers and publications, data, and software. Similar resources for their respective areas are made available (respectively) by the fuzzy sets and evolutionary computing portals.

WEB APPLICATIONS

Although to date, GC is mostly an area approached theoretically, there are important applications that recognize and exploit the knowledge that is present in data and therefore are impacted by GC developments. Such applications include data mining; data clustering, classification and aggregation; modeling of cognitive processes; and ontology learning. These become particularly relevant in the context of the Web, where data clearly differs from traditional text corpora in content, size, variability, and characteristics mentioned earlier (i.e., it is imprecise, uncertain, incomplete, or partially true). GC tools that are used for textual indexing and graph handling can be used for Web data mining.

DATA ANALYSIS

ROSETTA

http://rosetta.lcb.uu.se/

[analysis tool]

ROSETTA is a tool for analyzing tabular data based on rough set theory. It is designed with the objective to support the overall data mining and knowledge discovery process. It can perform tasks such as data preprocessing, browsing, computation of minimal attribute sets and generation of if-then rules or descriptive patterns, and validation of the induced rules or patterns. The tool and its source code are available for Windows and Linux platforms.

RSES

http://alfa.mimuw.edu.pl/~rses/

[analysis tool]

The Rough Set Exploration System (RSES) is a tool for analysis of tabular data based on rough set theory. The tool can calculate reducts (i.e., sufficient sets of features), generate decision rules using reducts, discretize numerical attributes, decompose large data into parts that share the same properties, search for patterns in data, manipulate data, and edit. The tool is available with a graphical user interface for Windows platforms.

DATA MINING AND CRAWLING

WEBGRAPH

http://webgraph.dsi.unimi.it/
http://law.dsi.unimi.it/

[graph framework]

The Web Graph from the University of Milan is a set of Java libraries for handling large graphs. It stores the data in compressed format and can achieve compression rates in the order of two to three bits per link by exploiting the regularity present in Web graphs. This makes working with Web data much faster as these graphs are typically huge in size. The library includes tools for creating symmetrical graphs and for transposing large graphs. A related library, available from the Laboratory for Web Algorithmics (LAW), includes code for efficiently executing the PageRank link analysis algorithm on these graphs. Both of these libraries are cross platform and available through the GNU Public License.

LUCENE

http://lucene.apache.org/
http://lucene.apache.org/nutch/
http://lucene.apache.org/hadoop/

[search and indexing tools, crawler]

Lucene is a platform centered on a Java-based text indexing and search system. It includes Nutch, a crawler component that can be used to build a Web search system, and Hadoop, a distributed platform for storing and processing data on a distributed file system. These are cross platform with Apache license.

TERRIER

http://ir.dcs.gla.ac.uk/terrier/

[search and indexing tool]

Terrier (TERabyte RetrIEveR) is a text search engine providing indexing and search capabilities over large corpora. It is written in Java and implements a set of features that are sufficient for providing a normal search engine right out of the box. Terrier is written with a focus on research so it is easy to extend and modify, for instance by providing new ranking or filtering methods for the search results. Terrier is cross platform with Mozilla Public license.

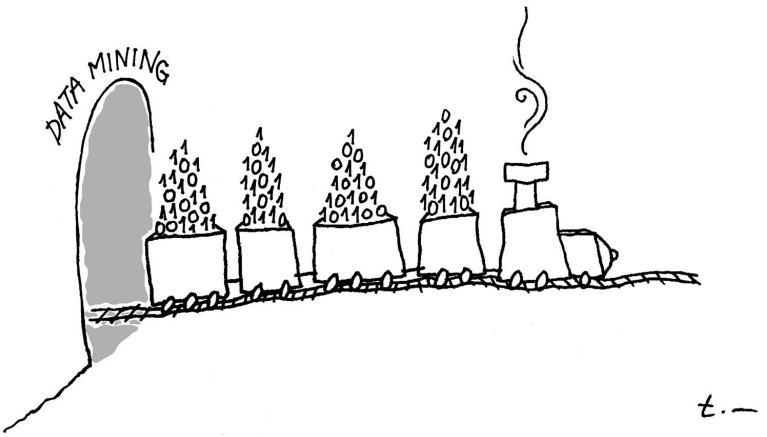
WEB DATA SETS

WEBSPAM-UK2006

http://www.yr-bcn.es/webspam/
http://webspam.lip6.fr/

[data set]

This is a large collection of over 11,000 hosts from the United Kingdom downloaded in 2006, with human-



Data mining. Cartoon by Tayfun Akgul (tayfun.akgul@ieee.org)

generated labels indicating if hosts are spam, nonspam, or suspicious. The collection includes a Web graph, a graph at the host level (hostgraph), the labels, and the contents of the pages. It is the collection used during 2007 for the Web Spam Challenges. This collection is freely available for research purposes.

WEB GRAPH DATA SETS

http://law.dsi.unimi.it/index.php?option=com_include&Itemid=65

[data set]

Several graphs from the Web are included in this page, including an 800K-nodes crawl, the Stanford WebBase-2001, and several crawls from Italy, the United Kingdom, and India between 2001 and 2005.

REUTERS 21578

http://www.daviddlewis.com/resources/testcollections/reuters21578/

[data set]

This is a data collection containing 21,578 stories published by Reuters in 1987, converted to the SGML format and annotated with topics by human editors. This data collection has been widely used for testing automatic text classification since 1996. The data collection is freely available for research purposes.

TREC

http://trec.nist.gov/

[data set]

TREC series of conferences on text retrieval (active since 1992) and their associated data sets have been mentioned in the context of digital video retrieval and multimodal signal processing in *IEEE Signal Processing Magazine* issues of September 2006 and March 2007, respectively. Each event consists of a series of tracks and used for testing competing solutions. Current tracks relevant to Web processing include question answering, e-mail spam filtering, blog exploration, enterprise search, etc. Interested teams must register with TREC to get the datasets and participate.

UCI KDD/ML

http://kdd.ics.uci.edu/

http://mlearn.ics.uci.edu/MLRepository.html

[data sets]

The two links above point to the University of California at Irvine Knowledge Discovery Archive and Machine Learning Repository. The former includes large data sets with a wide variety of application-specific and analysis task-specific data. The latter includes data sets, domain theories, and data generators that are widely used by the machine learning community for the empirical analysis of machine learning algorithms.

INEX

http://inex.is.informatik.uni-duisburg.de/

[data sets]

INEX contains several data sets that have been made available through a collaborative effort for studying content-based XML information retrieval. The goal in this effort is to build a reference data collection and propose metrics for the evaluation of XML retrieval systems. The databases are available to the INEX participants upon registration.

AUTHORS

Carlos Castillo (chato@yahoo-inc.com) is a researcher with Yahoo! Research in Barcelona, Spain.

Yiyu Yao (yyao@cs.uregina.ca) is a professor of Computer Science at the University of Regina, Canada.

life SCIENCES continued from page 123

ACKNOWLEDGMENT

This work was supported by NSF Grant 04-607 and by National Institutes of Health (NIH) Grant T15 HL088516.

AUTHORS

James B. Bassingthwaighte (jbb@bioeng.washington.edu) is a professor of bioengineering and radiology at the University of Washington. His research interests are in the areas of physiological transport and fractals in biology. He is a member of the U.S. National Academy of Engineering.

Howard Jay Chizeck (chizeck@ee.washington.edu) is a professor of electrical engineering and an adjunct professor of bioengineering at the University of Washington in Seattle. His research interests involve control engineering theory and the application of control engineering to biomedical and biologically inspired engineered systems. He is a Fellow of the IEEE.

REFERENCES

[1] J.B. Bassingthwaighte, "Strategies for the Physiome Project," *Ann. Biomed. Eng.*, vol. 28, no. 8, pp. 1043–1058, 2000.

[2] "Interagency opportunities in multi-scale modeling in biomedical, biological, and behavioral systems," National Science Foundation [Online]. Available: http://www.nsf.gov/pubs/2004/nsf04607/nsf04607.htm

[3] J.J. Saucerman, L.L. Brunton, A.P. Michailova, and A.D. McCulloch, "Modeling beta-adrenergic control of cardiac myocyte contractility in silico," *J. Biol. Chem.*, vol. 278, no. 48, pp. 47997–48003, 2003.

[4] B.H. Kuile and H.V. Westerhoff, "Transcriptome meets metabolome: Hierarchical and metabolic regulation of the glycolytic pathway," *FEBS Lett.*, vol. 500, no. 3, pp. 169–171, 2001.

[5] M.J. Herrgard, M.W. Covert, and B.O. Palsson, "Reconstruction of microbial transcriptional regulatory networks," *Curr. Opin. Biotechnol.*, vol. 15, no. 1, pp. 70–77, 2004.

[6] SBML: Systems Biology Markup Language [Online]. Available: http://sbml.org/index.psp

[7] CellML: Cell Modeling Markup Language [Online]. Available: http://www.cellml.org/

[8] P.J. Hunter, "Modeling human physiology: The IUPS/EMBS Physiome Project," *Proc. IEEE*, vol. 94, no. 4, pp. 678–691, 2006.

[9] JSim: Java-based Simulation Platform for Data Analysis [Online]. Available: http://www.physiome.org/jsim/

[10] A.C. Guyton, T.G. Coleman, and H.J. Granger, "Circulation: Overall regulation," *Ann. Rev. Physiol.*, vol. 34, pp. 13–46, 1972.

[11] S.R. Abram, B.L. Hodnett, R.L. Summers, T.G. Coleman, and R.L. Hester, "Quantitative circulatory physiology: An integrative mathematical model of human physiology for medical education," *Adv. Physiol. Educ.*, vol. 31, pp. 202–210, 2007.

[12] M.L. Neal and J.B. Bassingthwaighte, "Subject-specific model estimation of cardiac output and blood volume during hemorrhage," *Cardiovasc. Eng.*, vol. 7, pp. 96–119, 2007.

[13] J.-L. Coatrieux and J.B. Bassingthwaighte, *Proc. IEEE (Special Issue on the Physiome and Beyond)*, vol. 94, no. 4, pp. 671–677, 2006.

More resources are available at the Physiome Organization, http://www.physiome.org and http://www.physiome.org.nz.