AGENT TECHNOLOGIES FOR WEB MINING FROM A NON-EXTENSIVE THERMODYNAMICS PERSPECTIVE

Franciszek Grabowski, Marek Zarychta, Przemysław Hawro

***Abstract**: Motivated by the works that reveal the small-world effect and scale-free property of various real-life networks, many scientists devote themselves into studying complex networks. One of the ultimate goals is to understand how the topological structures of networks affect the dynamics upon them. In this paper, we give a brief review on the studies of agent technology appliances for Web mining which is performed in complex network environment ruled by non-extensive thermodynamics.*

***Keywords**: non-extensive thermodynamics; web mining; complex networks; multi-agent technology.*

***ACM Classification Keywords**: I.2.11 Distributed Artificial Intelligence - Multiagent systems*

Introduction

The study of complex systems has became one the most active areas of research. This subject has interest in natural sciences as well as in social and artificial systems. Typical features present in complex systems are long-range interactions, long-term memory, fractal phase-space structure, scale-free network structure, or even combinations of these characteristics. These systems are common in computer science and cannot be correctly described by the well-established Boltzmann-Gibbs statistics. Such systems lack linear relationship between the input and the output, or between a cause and its effect. Formally, such relationships are commonly described by a power law. Scientists are particularly interested in two types of dynamics in complex systems and complex networks:

- event-driven processes, a dynamic process is apparently caused by external interferences, for instance, a fire caused by a dropped match
- self-organized criticality, a system in which dramatic changes may take place in the absence of major causes at the macroscopic level [Chen, 2003].

Despite current complex network theory have not yet reached a level that one can specifically identify the cause–effect relations associated with the dynamics observed over a complex network, we should consider it as new approach to well known areas of research including web mining and multi-agent systems.

Network dynamics and topology of World Wide Web

Robust development in Information Technology and it's still gaining popularity has impact on the World Wide Web acting as an information super highway. Millions of people all around the world have access to miscellaneous information scattered over the Web. The distributed and dynamic nature of the Web has thrown down the gauntlet to information retrieval on this complex information space. The software agent technology seem to be appropriate to pick up this gauntlet. The system consistent of autonomous, collaborative and adoptive software entities is suitable for solving complex problems and coping with huge amount of information, but this software agents must be aware of their surrounding and run strictly connected to it.

Existing in real world networks, reveal more complex structures than were firstly considered, so the sense of topology evolved and acquired importance [Dorogovtsev, 2002]. Knowledge of topology is necessary to understand phenomena occuring in the network.

In the simplest random network model, vertex pairs are connected with each other with some probability value. Traditional approach to physical networks expose them as real finite dimensional regular lattices, field fully connected graphs or random graphs. Watts and Strogatz have developed model that interpolates between

a regular ordered graph and a random graph calling it small-world phenomenon [Watts, 1998]. Their model is appropriate for systems having a high degree of local clustering combined with a short finite path length between any pair of vertices. This feature is common among many social networks of human relationships, electric power grid, etc. A small-world network can be constructed starting from a regular lattice, in which each vertex is joined to its neighbours or fewer lattice spacings away, and then adding or moving a portion of the edges. The moving can be done by examining each vertex in turn and with some probability moving the other end of the edge to a different vertex chosen at random. The result is a lattice with shortcuts.

To study the degree distribution of a network, let p_k denote the fraction of vertices with degree k, or, the probability that a vertex chosen at random has degree k. Random network models produce usually a Poisson distribution for p_k , whereas most real-world networks have highly right-skewed degree distributions, meaning that there are lots of vertices having a few connections, and some vertices have many connections — highly-connected vertices are practically absent in random and small-world networks. The networks with right-skewed degree distributions have no characteristic scales for the degrees, thus the networks of this kind are called scale-free. Their degree distribution follows a power law $p_k \sim k^{-\gamma}$. Barabási and Albert built the model and found that the exponent in the power law has been approximated for many different real-world networks having values in range $2 < -\gamma < 4$ [Albert, 2002]. The World Wide Web topology based on preferential linking and revealing small-world phenomenon is currently one of the largest global social networks for which topological information is available.

Web mining

Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services [Dunham, 2006]. There can be seen huge grow and impressive evolution of the Web uncovering scalability problems of actual Web search engines. The Web structure can be compared to graph structure where pages appear as vertices and hyperlinks as edges. Collecting miscellaneous useful informations from this structure is called Web mining and can be divided into three categories:

1. Web content mining,
2. Web structure mining,
3. Web usage mining.

Web content mining aims at the knowledge discovery, where the objects are common collections of text documents, sets of multimedia documents such as images, videos, which are incorporated or linked to the Web pages. Web content mining from the agent-based approach aims on improving the information finding and filtering and could be divided into the following three categories [Cooley, 1997]:

- Intelligent Search Agents. These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.
- Information Filtering/ Categorization. These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.
- Personalized Web Agents. These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.

There is another approach to the Web content mining which aims on modeling the data on the Web into more structured form in the purpose of applying standard database querying mechanisms and data mining applications to analyze it. This approach is categorized to Multilevel databases and Web query systems.

Web structure mining concentrates on revealing the structure of the Web. The effort is focused on discovering the model underlying the hyperlink structures of the Web. This model can be used to categorize the Web pages and is useful to generate information such as similarity and relationships between Web sites containing important information which can help in filtering or ranking Web pages. Divergent nature of the Web and its objects creates new challenges and even obstacles, because there is not possibility to strictly make use of existing techniques such as database management or information retrieval. Since the Web is full of loops and traps, to generate a Web structure, the circuits and repetitive cycles should be detected and removed, without which a client may be

lost in Cyberspace through the complex cycles. Multi-agent systems seem to be well suited and have sufficient capabilities to perform such task.

Web usage mining covers the prediction techniques of the users behavior and their interactions with the WWW. The most useful source containing information for such analyses are logs from Web servers. The data collected from Web log records allows to discover user access patterns of Web pages and their interests. Typical applications generated from this analysis can be classified as personalization, system improvement, site modification, business intelligence and usage characterization [Cooley, 1997]. Web usage mining sometimes encounter obstacles for instance, due to the collaboration lack of the users or webserver administrators, who tend to keep the Web log records as jealously-guarded secret. Due to this fact, privacy plays a key role in Web usage mining, because users should be at least warned about privacy policies before they admit to reveal their personal data. There are no straight boundaries between Web structure mining and Web usage mining, hence all of them could be used in combined applications, which will not be discussed here.

Non-extensive thermodynamics and its impact on web mining techniques.

Since most of the systems in nature are in nonequilibrium states, not even tending toward equilibrium, over the years considerable effort has been devoted to the development of nonequilibrium statistical mechanics and nonequilibrium thermodynamics. The fluxes of energy, matter, etc. in nature are irreversible [Prigogine, 1961]. The second law of thermodynamics fixes the direction of these irreversible processes by specifying that the accompanying entropy production should be always positive. This is so indifferent of whether the system is open or closed, and independently of whether entropy flows into or out of the system to its surroundings. For an isolated system the second law therefore indicates that entropy can never decrease, but it does not affect open systems, where entropy can either increase or decrease. The Shannon information entropy defines entropy in terms of the probability distribution of observations, or the information applying to a set of observations. In nonequlibrium systems, irreversible behavior occurs as information is lost in the observation process. Driving a system away from equilibrium breaks symmetry and consequent emergence of organization but can also affect stability of the system by producing turbulences. Questions related to the statistical thermodynamics of irreversibility and self-organization are important in a wide range of new and cross-disciplinary fields, such as epidemiology, evolutionary dynamics, artificial life, agent technologies and web mining [Dewar, 2003].

Most systems, including Web, are not isolated they are open in various ways and they experience fluxes of energy, mass, information, etc. across their boundaries. Although the total entropy of a system plus its environment must increase, it does not follow that the entropy of an open system must increase. Instead, there are many remarkable instances of self-organization of such systems into coherent structures, ranging from tropical cyclones through individual biological organisms to human civilizations. Constantino Tsallis proposed new entropy definition to cope with all these phenomena [Tsallis, 1998]. So the question is not: whether self-organization of the Web will occur or not, but when it will occur and what are general principles to determine it?

In the last few years, there has been a great interest in understanding the topological properties of multi-mesh peer-to-peer networks which are capable of rearranging topology in case of node or link failure. This studies help us understand the behavior of systems such as the Internet and the World Wide Web. Whilst studies driven by traditional approach assume that once a link is created between two nodes, it is never deleted, research devoted to dynamic communication networks show that links are being constantly rewired. An important issue is to discover the topology that, given a search algorithm, optimizes the search process, optimality is defined as the minimization of the average time to perform a search. Clearly, being able to obtain such topology structures seems to be a useful guide to drive the evolution of dynamic communication networks [Cholvi, 2005].

Conclusion

Table1. Reductionism vs. System-Oriented Perspective.

	Reductionistic, Object-Oriented Approach	System-Oriented Approach
Principle	Behavior of a system can be explained the properties of its constituent parts	Multipartite systems have emergent properties that are only possesed by whole system and not the isolated parts
Model characteristics	Linear, predictable, deterministic	Non-linear, stochastic, sensitive to initial conditions, chaotic
Agent interactions	Homeostasis principle, normality, cooperation, self-regulation	Adaptation, robustness, percolations, self-organization
Network topology	Erdös-Rényi graphs, random linking, static wiring	Small-worlds, preferential linking, scale-free networks, dynamic, self-organizing topologies
Probability distribution	Poissonian or Gaussian distribution	Heavy tailed, self-similar, Pareto, Zipf distributions
Entropy definition	Classical Boltzman-Gibbs, Shannon	Tsallis non-extensive entropy
Information flows	Laminar, deterministic fluxes	Turbulent, self-similar fluxes

Still evolving network of hyperlinks and Web pages needs integration of different mining methods taking advantage of multi-agent architecture to permit the discovered knowledge to be verified, reliable and updated automatically. Automatic and non-invasive web personalization seems to be a challenge for nowadays search engines. As we can see in the table 1, underlying principles of web mining techniques should be revised in order for a system perspective to be fully appreciated. The reductionism nature of current paradigm manifests in many aspects leading to wrong perception of the Web. The developing fields of chaos theory, non-extensive thermodynamics and complex system science has not yet sufficient contribution to the Web mining. What becomes evident from these analyses is that the behavior of the system arises from the active interactions of its components appreciating emergent phenomena of coexistent entities. As it was said, the Web is kind of rapidly changing, uncertain environment with indeterministic information, thus for better understanding the phenomena occurring there, we need non-equilibrium thermodynamics and non-extensive statistics, which brings us adequate description of the stability in the open, distributed system in the state far from the thermodynamic equilibrium. Multi-agent systems are naturally suited to run in uncertain environments, for example in networks where there may be a connection or node failure, or someone can sabotage calculation by sending an incorrect data. Multi-agent systems do not require a synchronized clock, they easily adapt to the environment dynamically increasing and decreasing consumed memory size and processor usage (without any impact on performed calculations). They also tolerate a stopovers and frequent delay of communication as well as are capable of taking advantage of heterogeneous environments. The World Wide Web is an interactive and dynamic network in which the properties of single Web site is contingent on its relationships to other sites. Thus Web mining is performed with high risk of mistakes due to cheats and abuse done by unreliable Web-users, -masters etc. Widely used web mining techniques are susceptible to such unfair tricks because have not shifted to system oriented perspective yet.

Bibliography

[Albert, 2002] R. Albert and A.-L. Barabási: Statistical mechanics of complex networks. Reviews of Modern Physics, 74(1), 2002.

[Chen, 2003] Ch. Chen, N. Lobo: Semantically Modified Diffusion Limited Aggregation for Visualizing Large-Scale Networks, IV, Seventh International Conference on Information Visualization (IV'03), 2003.

[Cholvi, 2005] V. Cholvi, V. Laderas, L. Lopez, A. Fernandez: Self-adapting network topologies in congested scenarios. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), Vol. 71, No. 3. (2005).

[Cooley, 1997] R. Cooley, B. Mobasher, J. Srivastava: Web mining: information and pattern discovery on the World Wide Web, Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference, 1997.

[Dewar, 2003] R. Dewar: Information theory explanation of the fluctuation theorem, maximum entropy production and self-organized criticality in non-equilibrium stationary states, J. Phys. A: Math. Gen. 36, 631 (2003)

[Dorogovtsev, 2002] S. N. Dorogovtsev and J. F. F. Mendes: Evolution of networks. Advances in Physics, 51(4), 2002.

[Dunham, 2006] M. H. Dunham and S. Sridhar: Data Mining, Introduction and Advanced Topics, Prentice Hall Publication, 2006.

[Prigogine, 1961] I.Prigogine: Thermodynamics of Irreversible Processes, New York Interscience, 1961

[Tsallis, 1998] C. Tsallis: Possible generalization of Boltzmann-Gibbs statistics, Journal of Statistical Physics 52, 1998.

[Watts, 1998] D.J. Watts and S.H. Strogatz: Collective dynamics of „small-world" networks, Nature, 393, 1998.

Authors' Information

Franciszek Grabowski *– Editor, Department of Distributed Systems FECE RUT; W. Pola 2, 35-959 Rzeszów, Poland ; e-mail: fgrab@prz.edu.pl*

Marek Zarychta *– Editor, Higher Vocational State School in Jarosław; Czarnieckiego 16, 37-500 Jarosław, Poland ; e-mail:zarychta@pwszjar.edu.pl*

Przemysław Hawro *– Editor, Higher Vocational State School in Jarosław; Czarnieckiego 16, 37-500 Jarosław, Poland ; e-mail: przemyslaw@pwszjar.edu.*