Discovery of Knowledge Flow in Science

By Hai Zhuge

Recognizing and understanding knowledge flow between scientists is valuable for science. Discovering, managing, and utilizing such knowledge are advanced services of the e-science knowledge grid environment.

Whether we are aware of it or not, knowledge flows within human society and in the Internet-mediated interconnection environment scientists increasingly rely on for research. Knowledge flows influence the evolution of culture and language, promote international collaboration, and hasten the development of science.

Scientists have developed many approaches to the static representation of knowledge, and to extracting, discovering, learning, and reasoning about it. However, knowledge is dynamic—it goes through human brains for knowing, invention, propagation, fusion, generalization, and problem solving. Scientific articles are the major medium that carries knowledge between scientists.

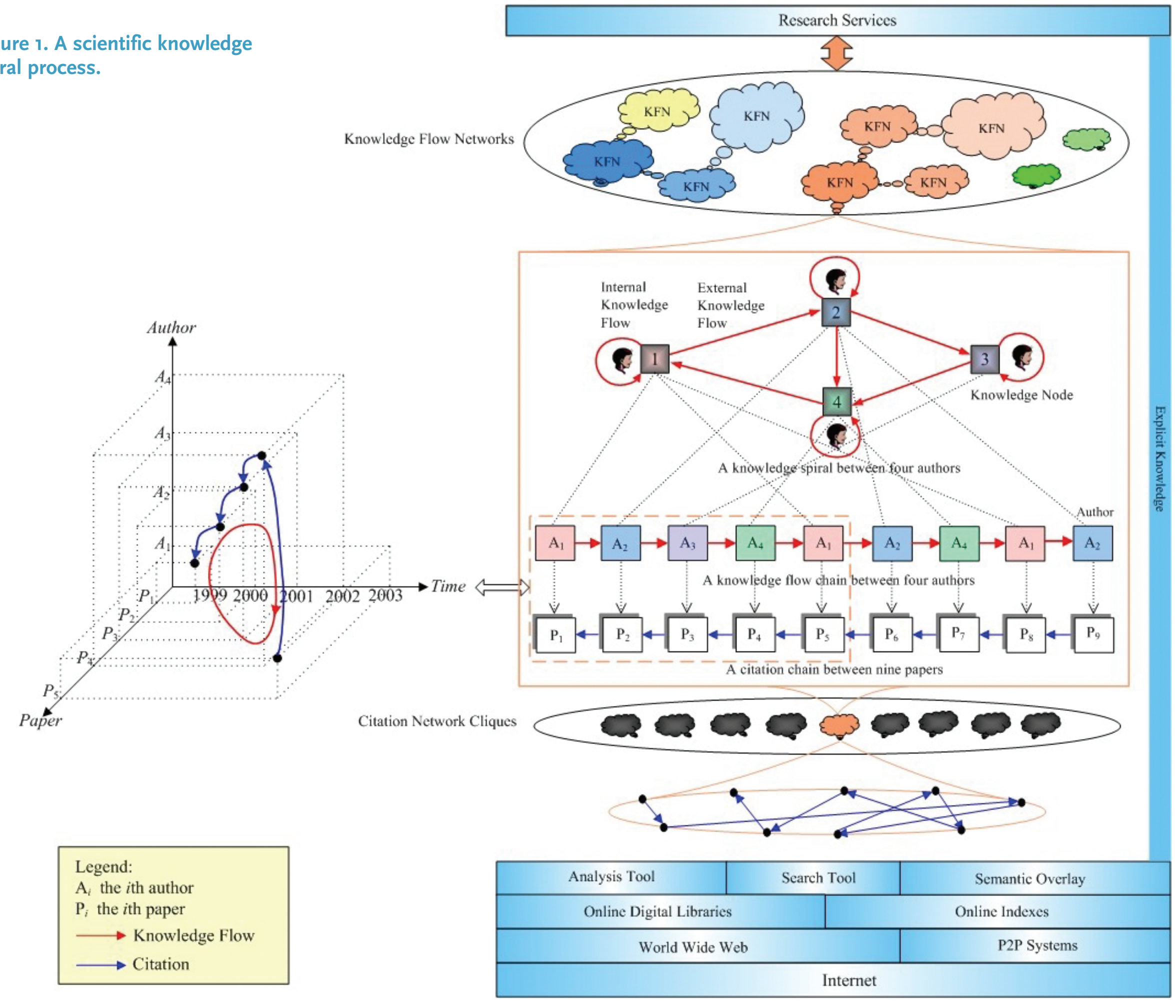
The Citation Network. The citations in scientific articles are objective data used, for example, by the Institute for Scientific Information's *Journal Citation Report*, for assessing scientists, scientific articles, and journals [5]. Statistics from ISI shows that top scientists are always most-cited.

Scientific articles' citations follow a power-law distribution [9], which is coincidentally similar to the distribution of the hyperlinks of the Web [1, 2]. Neither the citation nor the hyperlink relationship is transitive; that is, "A cites/links to B" and "B cites/links to C" does not imply "A cites/links to C."

Hyperlinking is arbitrary—"anything can link to anything" [3]—but citation is not. Citations among scientific documents form a *time-constrained non-redundant content net* with the following characteristics:

- The content of published papers is fixed.
- Citations between published papers are fixed.
- Published articles cannot cite those not yet completed.
- No two scientific papers may be completely identical.
- Authors who cited each other share some

Figure 1. A scientific knowledge spiral process.



knowledge, as do co-authors of a paper.

- The formation of an area is the process of forming a relevant citation community and its most-cited articles, which represent the area.

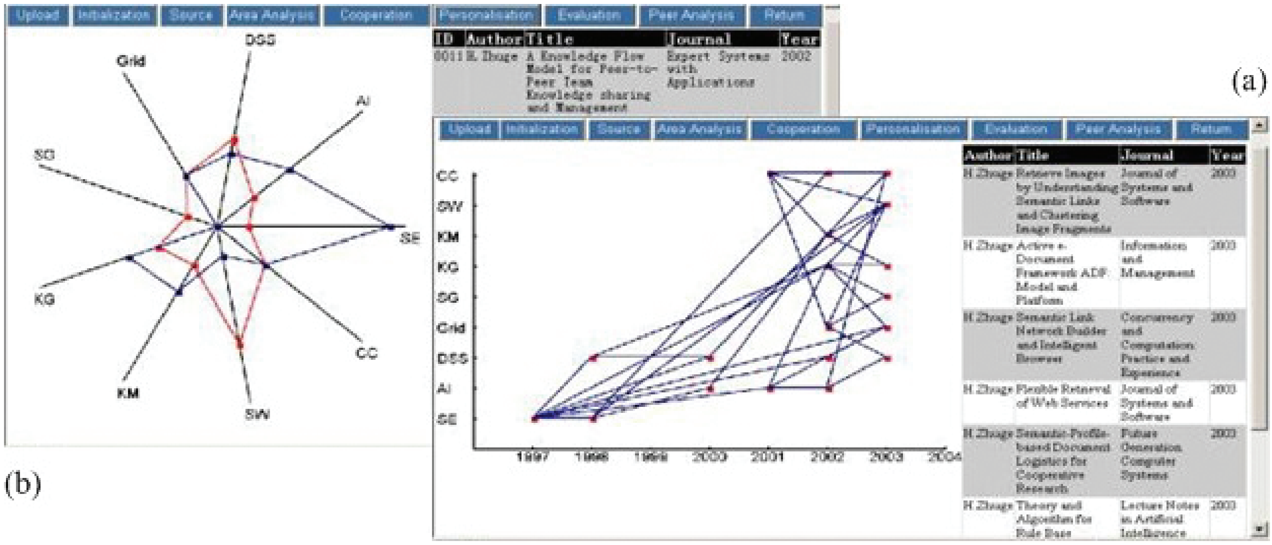
Knowledge Flows through a Citation Network

The ideas in a scientific article inspire new ideas, which will be recorded and published as new articles after peer review. Citations between scientific articles imply a knowledge flow from the authors of the article being cited to the authors of the articles that cite it.

The knowledge flow network implicit in the citation network consists of knowledge flows between nodes that process knowledge, including reasoning, fusing, generalizing, inventing, and problem solving, by authors and co-authors. As a scientific research area evolves, the knowledge flow network evolves with the citation network and behaves differently during different phases.

Roles	Definition	Explanation/Phenomenon
Source	A role that has published a small number of highly cited articles that cite not many other articles. A journal plays that role is a source journal. People play that role as a source scientist.	A source scientist published top highly cited articles of an area, their knowledge flow through citations network. A source outputs more knowledge than input. A source journal helps its articles to get rich in citations, which in turn helps the journal get higher impact. Scientists prefer to publish their articles in source journals.
Authority	An often-cited role that has published and cited a large number of articles.	An authority scientist contributes knowledge to the area but occupies others' growth space. Publishing review or survey articles helps a journal to be an authority journal.
Bee	An occasionally cited role who frequently publishes articles that cite diverse sources and authorities.	A bee has interest in many topics. It helps fuse, expand, explain and propagate knowledge. People play this role.
Hub	A seldom-cited role that has published and cited a large number of articles.	Hubs contribute to an area by reviewing articles, expanding and spreading knowledge. Journals and people play this role.
Novice	A role that has just started to publish a small number of articles that are seldom cited.	Novices contribute to an area by reviewing articles, bring new knowledge from other areas, spreading knowledge, and attracting new comers in their social networks. People and journals play this role.

Roles contributed to the development of an area. Scientists, articles, and journals play and change their roles with the evolution of an area.



ID	Author	Title	Journal	Year
0011	H. Zhuge	A Knowledge Flow Model for Peer-to-Peer Team Knowledge sharing and Management	Expert Systems with Applications	2002

Author	Title	Journal	Year
H. Zhuge	Retrieve Images by Understanding Semantic Links and Clustering Image Fragments	Journal of Systems and Software	2003
H. Zhuge	Active e-Document Framework ADF: Model and Platform	Information and Management	2003
H. Zhuge	Semantic Link Network Builder and Intelligent Browser	Concurrency and Computation: Practice and Experience	2003
H. Zhuge	Flexible Retrieval of Web Services	Journal of Systems and Software	2003
H. Zhuge	Semantic-Profile-based Document Logistics for Cooperative Research	Future Generation Computer Systems	2003
H. Zhuge	Theory and Algorithm for Rule Base	Lecture Notes in Artificial Intelligence	2003

Figure 2. Interface. (a) Depicting a citation network to show the evolution of interests and knowledge. (b) Displaying the distribution of cited and published articles.

Knowledge flows through the citation network without constraint. Scientists can publish in several areas, and thus be involved in different knowledge flow networks. This enables knowledge to flow through knowledge networks in different areas to promote interdisciplinary research.

An important characteristic of knowledge flow is its reachability—knowledge of author A's article can reach C when C cites B's article and B cites A's article.

The development of a large knowledge flow network involves certain special roles shown in the accompanying table. Different roles contribute differently to the development of an area. A researcher can play different roles in several areas.

The Knowledge Flow Spiral—A Knowledge Hypercycle Model

Knowledge flow spirals are formed when knowledge flows in a network. A knowledge node (scientist) can deliver knowledge to its peers by forwarding knowledge it has received (for example, forwarding an answer to the node that forwarded the query), or by passing on knowledge it generates (for example, send the answer to the querying node directly). The received knowledge inspires a node to generate new knowledge. Knowledge passing can take the form of broadcasting or query routing. Figure 1 depicts a knowledge spiral, comprising nodes and two types of flow:

- *External knowledge flow*—knowledge flowing between nodes, and
- *Internal knowledge flow*—knowledge arising within a node as the result of processing.

The functioning of the knowledge spiral is very similar to that of the hypercycle model [8]. The self-replication and catalytic-support arcs in the hypercycle correspond to the external knowledge flow and internal knowledge flow of the knowledge spiral. There are two differences: self-replication happens in the nodes of hypercycles while external knowledge flow is between nodes of a knowledge flow spiral, and catalytic support happens between nodes while internal knowledge flow is within individual nodes of a knowledge flow spiral. A knowledge flow spiral can generate and generalize knowledge during the processing and recycling of

The knowledge flow network implicit in the citation network consists of KNOWLEDGE FLOWS BETWEEN NODES THAT PROCESS KNOWLEDGE, including reasoning, fusing, generalizing, inventing, and problem solving, by authors and co-authors.

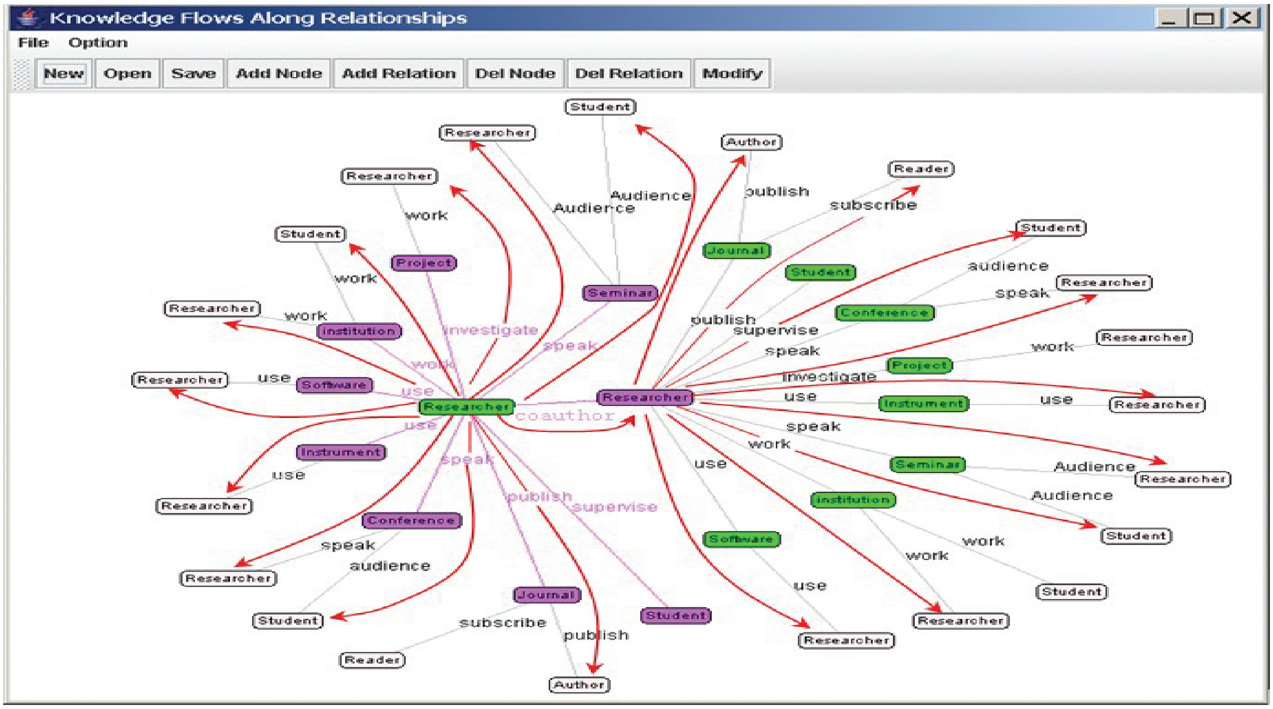


Figure 3. A visual semantic map organized according to interest. The red arcs denote knowledge flowing along the semantic links between scientists and scientific activities.

knowledge. An effective knowledge flow network enables a research team to be very powerful in generating new knowledge if every node can innovate and output new knowledge to appropriate members.

Recording the evolution of knowledge flows among authors needs a semantic space of three dimensions: *article, author*, and *time* as shown in Figure 1. An orthogonal classification semantic space called the Resource Space Model can ensure the correctness of operations on the space [10]. Storing knowledge requires a semantic space with three dimensions: *knowledge level, area,* and *location*. *Concept, axiom, rule, method,* and *theory* are coordinates of the knowledge level. Invisible knowledge flow networks self-organize and evolve continuously and interact with each other through flows to form an autonomous knowledge space for providing advanced knowledge services.

Knowledge energy and strategy of cooperation. Like energy in the physical world, knowledge energy reflects knowledge differences between nodes. A high-energy knowledge node in science has many articles being cited by peers, that is, it emits more knowledge and contributes more to an area. Reflecting the development of a discipline, the distribution of knowledge energy in a network changes as its citation network evolves. Maintaining knowledge energy differences between knowledge nodes and ensuring that only needed knowledge is passed between nodes are critical to realizing an effective knowledge flow spiral. Maintaining openness is a strategy for the sustainable development of knowledge flow within a research community.

The hyperlink network evolves under the "rich get richer" rule [6]—higher-ranking nodes have more energy to attract new links. For the same reason, highly cited articles attract more citations as the citation network evolves. However, the knowledge flow network evolves in a more complex and dynamic way. A high-energy knowledge node puts out more knowledge than low-energy nodes. However, the release of knowledge does not lead to the loss of energy: knowledge energy does not obey the physical world's law of the conservation of energy.

Knowledge energy can differentiate the capability of intelligent agents in large autonomous networks. Enhancing inflow, that is, selecting the appropriate higher-energy nodes to cooperate, is a strategy to increase the efficiency of knowledge flow networking. Our experiment shows that knowledge routing in a peer-to-peer autonomous network can attain higher efficiency if higher energy nodes are more likely to be selected at each hop.

Trusted team members ensure the effectiveness of knowledge sharing. In a trusted community, high-energy nodes that keep helping others will obtain

ENCOURAGING UNSELFISH COOPERATION is another stategy that helps a knowledge flow network reach its greatest effectiveness.

indirect help [7]. Everyone can benefit from the community if every node can generously contribute knowledge. Encouraging unselfish cooperation is another strategy that helps a knowledge flow network reach its greatest effectiveness.

Special knowledge spirals and their effect in science. Distinguishing the following special knowledge flow spirals can help scientists explore knowledge evolution in research community development:

- A *rising* spiral has an increasing rate of citations over time. This implies a rising research group or community. In contrast, a *descending spiral* has a decreasing rate of citations. This implies a declining research group or community.
- A *rising and expanding spiral* is a rising spiral that includes an increasing number of authors over time. This implies a rising and expanding research group or community.
- A *falling and shrinking spiral* is a descending spiral that is losing authors. This implies a declining and shrinking group or community.
- An *authoritative* spiral requires that all its nodes remain authoritative. This implies an authoritative research group.
- An *original* spiral has at least one source node. This implies the presence of an initiator.
- A *downstream* spiral contains authors whose articles often cite others but are seldom cited by other researchers.

USING AND MANAGING KNOWLEDGE FLOW NETWORKS IN E-SCIENCE

Scientists need an ideal e-science environment to support research more effectively than the Web. Scientists only need to start the environment by uploading their articles, or start with a directory or online database of articles. Searching the citation network among Web pages and in digital libraries can reveal knowledge flows among authors. By extracting and mining from scientific data, activities, and documents of an area; analyzing the relationships between results to enrich a knowledge flow network; and tracing its evolution, an e-science environment can provide scientists with the following services:

- *Outline personal research roadmaps and depict the evolution of interest and knowledge.* This service helps scientists record and analyze personal research history and status. They can use this information to plan their research by looking at the distribution and evolution of their areas. The system can display users' personal knowledge flow networks by finding their citations and their co-authors' citations, classifying the articles involved by discipline and chronology, expressing the category semantics for advanced services, linking the categories with the citations, and revealing knowledge flows using the citation network.
- *Recommend a network of appropriate references.* This service can recommend references by retrieving documents from the Web and digital libraries, ranking them according to their citation rates and the roles of the authors, and tracing and analyzing their citation networks to show the references as a network rather than as a list. The network can be a hierarchy that enables people to zoom in and out of the research area.
- *Display closely related peers and their status.* This service can help scientists select peers by finding the dense cliques in knowledge flow networks by looking at the statistical results and distribution of authors, articles, and citations.
- *Automatically discover interest groups.* This service can discover global interest groups by finding two types of citation relationship. The first relationship is between authors who cited the same article, on the basis that authors who cite the same article have a shared interest. The second is authors who cited each other; the more articles the authors mutually cite, the more common interests they share. Mutual citation can be extended to be indirect: if there is a citation path in each direction between two articles, then the authors share common interest and knowledge.
- *Estimate the effectiveness of cooperative research.* This service can help a researcher find prospective partners by their current roles (source, authority, bee, hub, or novice).
- *Objectively and dynamically evaluate research.* Criteria such as the number of citations, the number of articles, and the impact factor for journals are

used to evaluate research. Scientists can further find out who initiated a knowledge flow in an area, their roles in a flow, changes in their status, the effect of knowledge flows on different areas, and the evolution of a knowledge flow. The system can evaluate researchers, institutions, and disciplines.
- *Estimate the development stage of a discipline and its maturity.* Based on the evolution of relevant knowledge flow networks, this service enables scientists to simulate and estimate the development of a discipline, and thus helps them plan research.
- *Detect an emergent research area.* This service enables scientists to detect the emergence of a new area from the convergence of two areas, especially a mature area and a relevant new area.
- *Interact with scientists.*
  - Display the distribution of research communities and cooperation between institutions and between scientists by searching scientists' and their team's Web sites, and extracting co-authors, authors' affiliation and region, and sponsorship information from articles.
  - Inform scientists of the status and type of an interest or group of common interests as well as the evolution of a research area.
  - Answer queries about the research situation, including the number and distribution of researchers, articles, and topics over time.
  - Plan personal development, and make policies for the development of disciplines.

Current Web and text analysis techniques can automatically find the citation, co-author, and author-affiliation relations in scientific documents on the Web or in digital libraries. Extracting document fragments that contain authors, affiliations, and references, an escience knowledge grid environment can find useful semantic relationships between documents or document fragments. Analyzing the relationship between authors, articles, and their citation network, the environment can automatically discover knowledge flow networks and can distinguish the type of knowledge nodes using ranking algorithms and trace their evolution. By extracting comments and comparisons of the cited articles and then organizing them according to the time of publication, discovers and displays how the area has developed as a research literature.

Figure 2(a) shows the interface of the personalization function that shows the evolution of the user's interests and knowledge. The top portion is a list of operational functions, the left portion displays the citation network of the user's publications, and the right portion displays the list of publications of any network node that the user clicks. Figure 2(b) shows the interface of the other personalization function. The left portion displays the distribution over areas of the number of publications and of citations using color-coded curves. The user can click any point to display relevant publications in the right portion.

Knowledge also flows in scientific activities such as communication between collaborators. The environment records information flow such as email within a research team, extracts useful information flows, analyzes common interests and cooperation relations, and displays the evolution of the flows to help assess the teamwork's effectiveness, for example, whether team members' changing interests match their tasks.

Knowledge Flows through Semantic Link Networks

Semantic links exist between scientists, scientific activities, and scientific entities such as journals and research institutions. These semantic links constitute a scientific semantic map [10]. Figure 3 shows a tool for visualizing the semantic map where the center nodes can trace the user's interest. Knowledge flows along semantic links such as "co-author" and "supervise" prior to other links to constitute a knowledge map. Such a knowledge map is dynamic, and it could be discovered by analysis of these links.

Research has shown the contact network and the virus spread model determine the spread of epidemics [11], so the evolution of the contact network influences an epidemic. Appropriately changing the contact network can control an epidemic. Knowledge

Coordinating and fusing knowledge flows, data flows, and control flows, and integrating knowledge flows and workflows, are POWERFUL MEANS FOR MAKING TEAMWORK EFFECTIVE.

flows along a semantic link network by such models as knowledge dissemination and query routing. Changing the semantic link network will influence the efficiency of query routing. The knowledge dissemination model resembles the spread of infectious diseases. Changing the semantic link network will influence the knowledge flowing through it.

Unlike data and control flows in workflows, which are predefined or adapted according to activities and changes, knowledge flows are behind the interaction among team members and come with activity-level cooperation. They work on knowledge-level cooperation.

Knowledge flows fuse frequently when used, and their contents are not predictable. Effectively managing knowledge flows within a team can lead to effective team knowledge management and eventually raise the effectiveness of teamwork. Coordinating and fusing knowledge flows, data flows, and control flows, and integrating knowledge flows and workflows, are powerful means for making teamwork effective.

Knowledge flow in future interconnection environments. Computing professionals are striving to develop new computing and interconnection platforms to provide more advanced services [3, 4]. The future interconnection environment will be a large-scale human-machine environment where the physical world, mental world, and digital virtual world will interact and evolve cooperatively [12]. Knowledge flow will be the major engine of its evolution. Various knowledge flow networks interact with each other and evolve continuously to constitute the dynamic overlay of the autonomous knowledge grid environment [10].

The development of science depends on scientists and on their recording and sharing ideas. The future interconnection environment will synergize data flows, control flows, and knowledge flows to support appropriate on-demand service flows [11, 12], from which knowledge flow networks will create a live knowledge space that develops independently of machines. Interacting with scientists and their records, this environment will evolve in a novel way to help the development of science.

Conclusion

Exploring the universe and human society are great challenges of 21st century science. This article explores the dynamic nature of knowledge, the power to promote and influence the development of human society, and the future interconnection environment. It describes an important approach to automatically discovering knowledge flow networks within scientific documents and activities. Such networks embody an autonomous knowledge grid, which supports individual and cooperative scientific research, helps investigate the evolution of knowledge and disciplines, and assists in planning for scientific research development. Using the rules of knowledge flow makes teamwork more effective and innovative, not only in an e-science environment but also in team management, as knowledge flow networking can generate knowledge during operation. However, many other scientific and technological issues relevant to knowledge flow, including human, philosophical, psychological, cognitive, economical, social, and management issues, are yet to be resolved, and are a major challenge to scientists. C

References

1. Adamic, L.A. and Huberman, B.A. Power-law distribution of the World Wide Web. *Science, 287,* 24 (2000), 2115.
2. Barabási, A.L. and Albert, R. Emergence of scaling in random networks. *Science, 286* (1999), 509–512.
3. Berners-Lee, T., Hendler, J., and Lassila, O. Semantic Web. *Sci. Am. 284,* 5 (2001), 34–43.
4. Foster, I. Service-oriented science. *Science 308,* 5723 (2005), 814–817.
5. Katerattanakul, P., Han, B., and Hong, S. Objective quality ranking of computing journals. *Commun. ACM 46,* 10 (2003), 111–114.
6. Kleinberg, J. and Lawrence, S. The structure of the Web. *Science 294,* 30 (2001), 1849–1850.
7. Nowak, M.A. and Sigmund, K. Evolution of indirect reciprocity. *Nature 427*, 27 (2005), 1291–1298.
8. Oida, K. The birth and death process of hypercycle spirals. *Artificial Life VIII,* R.K. Standish, M.A. Bedau, and H.A. Abbass, Eds. MIT Press, New York, 2002.
9. Redner, S. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J.* B4 (1998), 131–134.
10. Zhuge, H. *The Knowledge Grid.* World Scientific Publishing Co. Singapore, 2004.
11. Zhuge, H. Exploring an epidemic in an e-science environment. *Commun. ACM 48,* 9 (Sept. 2005), 109–114.
12. Zhuge, H. The future interconnection environment. *IEEE Comput. 38,* 4 (2005), 27–33.

Hai Zhuge (zhuge@ict.ac.cn) is a professor and the director of the Key Lab of Intelligent Information Processing of the Chinese Academy of Sciences, Beijing, China.

This research work was supported by the National Basic Research Program of China (973 project no. 2003CB317001) and the National Science Foundation of China.