Information Overload

Sonia Bergamaschi and Francesco Guerra
University of Modena and Reggio Emilia

Barry Leiba
Huawei Technologies

This is the Information Age — and information is everywhere. We once got our news from newspapers, for example. But with advances in technology, we gained new options, from radio to broadcast television to 24-hour news channels on cable and satellite. Today, we can read any newspaper in the world on the Internet. We have, essentially, infinite news sources. That's great, right? Or is it too much?

In "the old days," when someone called you on the telephone at work, that person could leave a message if you were busy or out. Back then, people avoided calling each other with trivial matters, so they generally opted to leave a message only if the call was vitally important. And you could talk only with one caller at a time. Today, we have email, and it's much easier for people to leave messages. It's also much easier for people to email each other about trivial matters and routinely copy everyone on every discussion, whether they need to be copied or not. Spam contributes to the problem, but even without spam, most people would agree that email is out of control. Add in blogs, Facebook, Twitter, online office bulletin boards, and corporate messaging and discussion systems, and we've clearly hit information overload.

What Is Information Overload?

Search the Internet for the phrase "information overload definition," and Google will return some 7,310,000 results (at the time of this writing). Bing gets 9,760,000 results for the same query. How is it possible for us to process that much data, to select the most interesting information sources, to summarize and combine different facets highlighted in the results, and to answer the questions we set out to ask? Information overload is present in everything we do on the Internet.

Despite the number of occurrences of the term on the Internet, peer-reviewed literature offers only a few accurate definitions of information overload. Among them, we prefer the one that defines it as the situation that "occurs for an individual when the information processing demands on time (Information Load, IL) to perform interactions and internal calculations exceed the

 1089-7801/10/$26.00 © 2010 IEEE

supply or capacity of time available (Information Processing Capacity, IPC) for such processing."[1] In other words, when the information available exceeds the user's ability to process it. This formal definition provides a measure that we can express algebraically as IL > IPC, offering a way for classifying and comparing the different situations in which the phenomenon occurs. But measuring IL and IPC is a complex task because they strictly depend on a set of factors involving both the individual and the information (such as the individual's skill), as well as the motivations and goals behind the information request.

Clay Shirky, who teaches at New York University, takes a different view, focusing on how we sift through the information that's available to us. We've long had access to "more reading material than you could finish in a lifetime," he says, and "there is no such thing as information overload, there's only filter failure."[2] But however we look at it, whether it's too much production or failure in filtering, it's a general and common problem, and information overload management requires the study and adoption of special user- and context-dependent solutions.

Due to the amount of information available that comes with no guarantee of importance, trust, or accuracy, the Internet's growth has inevitably amplified preexisting information overload issues. Newspapers, TV networks, and press agencies form an interesting example of overload producers: they collectively make available hundreds of thousands of partially overlapping news articles each day. This large quantity gives rise to information overload in a "spatial" dimension — news articles about the same subject are published in different newspapers — and in a "temporal" dimension — news articles about the same topic are published and updated many times in a short time period.

The effects of information overload include difficulty in making decisions due to time spent searching and processing information,[3] inability to select among multiple information sources providing information about the same topic,[4] and psychological issues concerning excessive interruptions generated by too many information sources.[5] To put it colloquially, this excess of information stresses Internet users out.

Achievements and Challenges

To manage information overload, a user is required to discriminate among useful, redundant, incorrect, and meaningless information. From a computer science perspective, this means we must provide users with a combination of techniques and tools for collecting, grouping, classifying, selecting, indexing, ranking, and filtering useful information. Generally speaking, all these processes share a common goal: they aim to match users' information needs with available information. In other words, managing information overload requires that we extract a semantic description from information sources, understand the meanings behind user requests, and match these semantic elements. The research community has been addressing issues related to these requirements, resulting in a growing collection of achievements.

Description of Information Sources

Information sources typically don't convey semantics that can be processed by automatic applications. Consequently, for software applications, knowing which contents a source holds is a complex task. Researchers have developed technologies in the Semantic Web (www.w3.org/standards/semanticweb/) and in database communities,[6] associating semantic knowledge with data sources. The main ingredients of these technologies are ontologies and vocabularies (allowing the association of metadata to data source elements), reasoners (allowing the discovery of mappings among different data sources), and matching and integration techniques (enabling the interoperability of different data sources).[7]

Sources frequently change their contents, so "on the fly" technologies are important in real applications. However, the use of such technologies isn't always possible, due to the computational load they require. On the other hand, with "offline" techniques, users don't know if they're dealing with updated data because information users and producers are, in general, different, and there's currently no standard mechanism for notification of source changes. RSS, Atom, and other syndication technologies address this issue. Nevertheless, they can be exploited effectively only by human users because they don't implement a standard protocol for conveying certain machine-understandable semantics.

Formulation of Users' Needs

A user generally has two main possibilities for acquiring information: queries expressed in a structured query language along with

keyword-enabled navigational searches (the user provides a combination of words which he expects to find in the document),[8] and research searches (the user provides a combination of words denoting an object about which the user is trying to gather/research information).

Query languages let users formulate complex queries, with selection clauses defining accurate constraints on the results. But they require that the user know the source's structure and contents (table names, attribute domains and values, and relationships among structures) and a language for formulating effective queries. The research community has been involved in developing tools for supporting users in writing queries (see the query-by-example approach,[9] for one) and for visualizing data-source structures.[10]

Keyword-based searching is a way of dealing with some of the previously mentioned issues here. A keyword query is a list of keywords that are related in some way to the elements of interest. The list is typically flat: no relationship is provided among the keywords, and there are no dependencies, no semantics about each keyword, and no information about the roles they play in relation to the data repository structures.

Keyword-based searching has been extensively studied and used in the area of information retrieval. It typically works by building a set of specialized indexes over sets of documents and then using these indexes to identify the documents that contain as many of the keywords provided in a query as possible.[11] One of the main problems in this area is to discover the meaning intended by a user formulating a keyword query — a short list of keywords doesn't provide enough context for disambiguating the terms. The same issue happens in the social network scenario, in which the retrieval of specific contents is a complex task, mainly because of the lack of semantics in tag definition.[12,13]

Matching Users' Needs and Available Data

Filtering, recommendation, and searching processes support users in discerning the messages that satisfy personalized criteria among the ones received. The main issues in this area concern how to index and manage source descriptions; the development of semantic, approximate, collaborative techniques for matching user requests and available data;[14] the selection of the join paths connecting the relevant data structures intra- and intersource; and the use of users' feedback for improving results.[15]

In This Issue

Due to the amount of information available, one of the most interesting strategies for addressing information overload is determining relevance measures for selecting or retrieving information that best matches user requests. Nevertheless, selecting which information is useful for a user is a critical task because incorrect and incomplete information can have consequences on significant decisions in both business and private sectors. This critical situation is even more manifest at the Internet level, where a few results provided by a few search engines are the entry points for billions of webpages. This asymmetry in information management might be the cause of a double risk. On one side, information producers might devote more attention to including tricks to get the first positions in a search engine (search engine optimization) than to the quality of the information conveyed. On the other side, because users generally rely on the first search results, several multifaceted information sources aren't considered. The risk is that all users have the same thoughts because they all take the same flat information into account.

The four articles selected for this issue try to address this intrinsic risk shared by the most search techniques. "Search Computing: Managing Complex Search Queries" describes the core concepts of search computing applications — that is, applications supporting (or replacing) users in their ability to decompose queries and manually assemble complete results from partial answers, each with its own ranking. A recommendation system exploiting information obtained from social tagging is proposed in "Exploiting Social Tagging in a Web 2.0 Recommender System." The remaining two articles each propose a solution for managing information overload in particular scenarios. "Addressing Information Overload in the Scientific Community" presents the "liquid journals" model, which aims to facilitate the search for (and navigation of) scientific information of interest via a collaborative filtering system; "Overcoming Information Overload in the Enterprise: The Active Approach" proposes an integrated knowledge management workspace that reduces information overload by analyzing

context and user behaviors, and enables information sharing via tagging, wikis, and ontologies.

These selected articles show four different facets of the information overload problem, providing the big picture of the main research areas. This is a challenging research topic that includes the personalization required by different domains, users, and purposes, the growing amount of information available (heterogeneous in quality, structure, content, language), and real-time requirements. The research community has been addressing information overload in the past decade, but several problems are still open. The aim of this special issue is to provide some insights into the most recent research activity.

References

1. A.G. Schick, L.A. Gordon, and S. Haka, "Information Overload: A Temporal Approach," *Accounting, Organizations and Society*, vol. 15, no. 3, 1990, pp. 199–220.
2. R. Juskalian, "Interview with Clay Shirky, Part I," *Columbia Journalism Rev.*, 19 Dec. 2008; www.cjr.org/overload/interview_with_clay_shirky_par.php?page=all.
3. A.F. Farhoomand and D.H. Drury, "Managerial Information Overload," *Comm. ACM*, vol. 45, no. 10, 2002, pp. 127–131.
4. X. Yang, C.M. Procopiuc, and D. Srivastava, "Summarizing Relational Databases," *Proc. Very Large Data Base Endowment*, vol. 2, no. 1, ACM Press, 2009, pp. 634–645.
5. E. Russell, L. Millward Purvis, and A. Banks, "Describing the Strategies Used for Dealing with Email Interruptions According to Different Situational Parameters," *Computers in Human Behavior*, vol. 23, no. 4, 2007, pp. 1820–1837.
6. A. Doan and A.Y. Halevy, "Semantic Integration Research in the Database Community: A Brief Survey," *AI Magazine*, vol. 26, no. 1, 2005, pp. 83–94.
7. S. Bergamaschi and A. Maurino, "Toward a Unified View of Data and Services," *Proc. 10th Int'l Conf. Web Information Systems Eng.* (WISE 09), LNCS 5802, Springer, 2009, pp. 11–12.
8. R.V. Guha, R. McCool, and E. Miller, "Semantic Search," *Proc. World Wide Web Conf.*, ACM, 2003, pp. 700–709.
9. M.M. Zloof, "Query by Example," *AFIPS Nat'l Computer Conf.*, AFIPS Press, 1975, pp. 431–438.
10. A. Katifori et al., "Ontology Visualization Methods: A Survey," *ACM Computing Surveys*, vol. 39, no. 4, 2007, article 10; http://portal.acm.org/citation.cfm?doid=1287620.1287621.
11. J. Zobel and A. Moffat, "Inverted Files for Text Search Engines," *ACM Computing Surveys*, vol. 38, no. 2, 2006, article 6; http://portal.acm.org/citation.cfm?doid=1132956.1132959.
12. J.G. Breslin and S. Decker, "The Future of Social Networks on the Internet: The Need for Semantics," *IEEE Internet Computing*, vol. 11, no. 6, 2007, pp. 86–90.
13. H.-L. Kim et al., "The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies," *Proc. 8th Int'l Conf. Dublin Core and Metadata Applications*, Dublin Core Metadata Initiative, Singapore, and Universitätsverlag Göttingen, 2008; http://edoc.hu-berlin.de/browsing/conferences/index.php.
14. F. Giunchiglia, M. Yatskevich, and P. Shvaiko, "Semantic Matching: Algorithms and Implementation," *J. Data Semantics*, vol. 9, LNCS 4601, Springer, 2007, pp. 1–38; doi: 10.1007/978-3-540-74987-5_1.
15. *IEEE Data Engineering Bulletin*, Special Issue on Keyword Search, vol. 33, no. 1, 2010, pp. 3–78.

Sonia Bergamaschi is a full professor of computer engineering at the University of Modena and Reggio Emilia. She's also director of the Information Engineering Department at the University of Modena and Reggio Emilia and leads the Database Research group (www.dbgroup.unimo.it). Her research activities include data integration systems, database technology, knowledge representation, reasoning techniques applied to databases, and the Semantic Web. Bergamaschi is a member of IEEE and the ACM. Contact her at sonia.bergamaschi@unimore.it.

Francesco Guerra is assistant professor of computer engineering at the University of Modena and Reggio Emilia, where he teaches enterprise information systems. His main research interests include integration of heterogeneous information sources, keyword search in structured databases, ontologies, and the Semantic Web. Guerra has a PhD in information engineering from the University of Modena and Reggio Emilia. Contact him at francesco.guerra@unimore.it.

Barry Leiba is an Internet standards manager for Huawei Technologies. His research interests include email and related technology; antispam work, messaging, and collaboration on mobile platforms; security and privacy of Internet applications; and Internet standards development and deployment. Leiba chairs four working groups in the IETF, was program chair for the 2010 Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, and is the editor for the Standards column in *IEEE Internet Computing*. Contact him at barryleiba@computer.org.