Viegas, Evelyne, "Data as an Enabler of Open Innovation: Challenges and Opportunities"

Changing the Conduct of Science in the Information Age Workshop, November 12, 2010

Data as an Enabler of Open Innovation

Challenges and Opportunities

Evelyne Viegas

Microsoft Research

evelynev@microsoft.com

It has not become any easier to find a needle in a haystack in the information age.

With over 20 billion pages, images, video and audio files on the surface web, and growing, written in over 80 languages, delivered in different formats, and a deep web that has hardly been indexed, finding information is not becoming easier. To reach near 100% accuracy and to transform data into knowledge relevant to the information seeker, we need to go beyond string manipulation and towards semantic data to better support information discovery and enable decision making.

Data needs to be transformed in information and knowledge in the context of a user or situation and needs to be accessible by anyone, from anywhere, at anytime.

To enable the paradigm data, information, knowledge, intelligence, much research and innovation are still needed in various areas including technical, sociological, legal, economical and societal. From a technical viewpoint, this means that researchers need to have access to real world large scale data, some of which that cannot be made available without restriction due to privacy and proprietary sensitivities. We focus below on three areas which present, in our opinion, real opportunities to accelerate research while calling for some changes in the way research is performed.

Cloud computing to address data overload – Cloud computing is evolving into a prerequisite to developing applications due to the amount of data out there and data compute to process it: multimedia data, social networks, computer vision. It is becoming more and more difficult to move data around or to compute on it locally to perform research, and new models to conduct data-driven research, such as cloud computing, are necessary. Being able to reproduce results is at the core of scientific endeavors. However, today scientists by working on obsolete data benchmarks they may be reproducing results of already obsolete trends. Should researchers focus on available data sets at the expense of large scale timely data which changes regularly and could be accessed via services? Of course such a proposition brings the challenges of defining new models to support reproducibility when the data itself cannot be shared or when the results are dependent on software when data is accessed via services. Is access to data,

© 2010 Microsoft Corporation. All rights reserved.

hard to find, enough of an incentive for a researcher to embrace new research models and evaluations?

Data access with privacy in mind – Some data cannot be made available for research because of some sensitivity attached to the data, such as for instance user logs which contain individual privacy or business assets which contain commercial value. And yet, this is data which may yield to societal discoveries if it could be analyzed. Data anonymization which allows performing research has proven difficult in practice with well documented privacy breaches. Privacy-preserving approaches may be helpful in some cases while being too restrictive to allow research in others (e.g. privacy-integrated queries). Another approach, in line with cloud computing, may be to leave the data securely hosted with the data owner, while allowing dynamic access to it via a query engine or service (e.g. www.research.microsoft.com/web-ngram which exposes n-grams probabilities to researchers based on the Bing search engine index). With such an approach, it becomes easier to provide recent and timely data to be accessed, but the notion of data benchmarks for science reproducibility can easily disappear. Should we focus on sharing data or should we focus on data access and finding new models to support science while accounting for the dynamicity of data?

Semantic Data for decision making – Data has become a 1st class citizen under different multimedia encoding: text, speech, non verbals, images, videos, sensors, and semantics is emerging as a unifying paradigm. In a knowledge-driven society the emergent ecosystem of software and services for research will require technologies which enable machine-based information management, analysis, reasoning, and inference. Products and tools from information industries are underway to start delivering on the promise of semantic computing (e.g. visual search, semantic search). However, we need further investment in and wider deployment of semantics-based technologies, such as those demonstrated by research projects funded by UK eScience and the NSF Cyber-enabled Discovery and Innovation programs, and which can now be scaled up to web-scale via the emergent cloud computing infrastructure and the availability of Linked Data.

© 2010 Microsoft Corporation. All rights reserved.