Data Externality

Rakesh Agrawal
Search Labs, Microsoft Research
1065 La Avenida
Mountain View, CA 94043
rakeshA@microsoft.com

ABSTRACT

In economics, an externality is an indirect effect of consumption or production activity on agents other than the originator of such activity. We observe that internet is enabling the design of information services that become smarter more they are used because of the data generated in the process. We give examples from web search to make the notion of data externality concrete and propose that thinking and designing for data externality could be an interesting direction for future data research.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data Mining*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Search Process*. H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.*

General Terms

Algorithms, Design, Economics, Human Factors

Keywords

Externality, Web Search, Health, Education

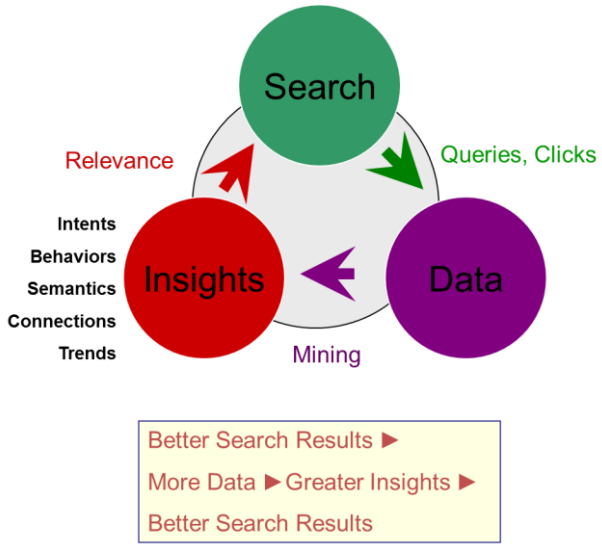
1. INTRODUCTION

Externalities are indirect effects of consumption or production activity, that is, effects on agents other than the originator of such activity which do not work through the price system [1]. A well-known example of positive externality is the network effect - an individual buying a device that is interconnected in a network increases the usefulness of similar devices to other people who already have them without having to pay extra for this increase in their utility function.

Internet is enabling the design of information services that become increasingly smarter more they are used by making use of the data which is generated as the users interact with the service. We call this positive spillover effect the *data externality*. However, every online information service does not automatically benefit from this effect. They need to be carefully engineered for them to exhibit positive data externality. Our view is that the data researchers can play pivotal role in the design and deployment of such information services.

In order to make the discussion concrete, we first discuss data externality in web search. Specifically, we describe two representative applications that have been designed with data externality in mind. We then consider two other areas – health and education – and suggest how data externality can be incorporated in these domains. We conclude by outlining some research opportunities that data externality offers to us data researchers.

2. SEARCH & DATA: VIRTUOUS CYCLE



There exists a strong virtuous cycle between web search and user data. As users pose search queries and click on search results, these queries and clicks are recorded by the search engine, which are then mined by the search engine to develop greater insights, which in turn help search engines provide better search results, leading to increased number of queries and clicks at the search engine. The greater the number of queries and clicks a search engines sees, the better it can get in producing more relevant results. We describe next two illustrative instances of this phenomenon.

2.1 Ranking

A key determinant of user satisfaction with search results is the quality of ranking. A popular way of ranking web results is by employing a learning algorithm [2]. The performance of a learning algorithm critically depends on the quantity and quality of training data [3]. The training data for this purpose consists of

5th Biennial Conference on Innovative Data Systems Research (CIDR '11) , January 9-12, 2011, Asilomar, California, USA.

tuples of the form <*q, d, l*> where the label *l* is the relevance judgment assigned to the document *d* for a query *q*. The judgment can be on a multi-point scale, ranging from perfect to bad. For each training tuple, certain number of query independent features (e.g. static page rank of the document) as well as query dependent features (e.g. position of a query word in the title of the document) are computed, which are then used for training the learning algorithm. But this begs the question – how are the <*q, d, l*> tuples generated in the first place?

Queries used in the training data are sampled from the query log. Clearly, the larger the number of queries a search engine sees, the more representative will be the sample. Trickier is the problem of obtaining labels for query-document pairs. Labels can be generated by employing human judges. However, obtaining a large number of judgments, particularly when one is dealing with documents in multiple languages can be expensive and time consuming. A more subtle issue is that the task of manual judging is not easy. Consider the query, *wink*, and two documents: one that is the home page of the Winks statistical software and another that contains various emoticons. How would you judge them, when you do not know what was the intent of the query?

The way out of this dilemma posited in [4] is to use the log of page impressions and documents clicked by various users for a given query and use this data for algorithmically generating labels. The exact details of the algorithm are not important for the purposes of this paper. The essential idea is to construct a per-query preference graph in which documents constitute vertices and weights on edges between vertices correspond to estimated number of users who prefer one document over the other based on click frequencies. Labels are then assigned to vertices by cutting the graph into as many partitions as the desired number of labels in such a way that the difference in the weights of edges that agree with the labeling and those that disagree is maximized, and thus maximizing user satisfaction. In the general case, when the number of label classes is unlimited, the problem is NP-hard. However, optimal partitioning can be obtained in linear time for the two class problem and effective heuristics exist for the general case.

Clearly, the key determinant of the effectiveness of the above procedure is the number of users using the search engine. As the number of users posing queries and clicking on search results increases, the coverage and the quality of the labels improve, leading to better ranker and better search results.

2.2 Diversity

A simple search box into which the user can type whatever string the user thinks best describes the information user is looking for has become a universal interface and the simplicity of this interface has been a key factor contributing to the immense popularity of web search. A consequence of this simplicity is that different users provide to the search engine the same query string even when they have very different information goals in mind. Joe might be looking for a hedge trimmer, whereas Dan might be looking for a beard trimmer, but both may ask the query – *trimmers*. While the users asking them have unique intents, the intent of many queries looks ambiguous to the search engine. How can then the search engine best answer such queries?

The approach taken in [5] is to diversify the search results so that the probability that an average user will find at least one relevant document in the retrieved result is maximized. Thus, a result page serves dual purpose - it offers relevant results for multiple intents of a given query and once the user has signaled a specific intent by a clicking on a document, further results take into account this disambiguation information.

The problem is formalized as follows. Given query *q*, a set of documents *D*, a probability distribution of categories for the query $P(c|q)$, the quality values of the documents $V(d|q,c)$, and an integer *k*, find a subset of documents *S* with $|S| = k$ that maximizes $P(S|q) = \sum_c P(c|q) (1 - \prod_{d \in S} (1 - V(d|q,c)))$. If $V(d|q,c)$ were interpreted as the probability that document *d* satisfies a user that issues query *q* with the intended category *c*, then the product term signifies the probability that the set of documents in category *c* will all fail to satisfy and one minus that product equals the probability that some document will satisfy category c. Finally, summing up over all categories, weighted by $P(c|q)$, gives the probability that the set of documents *S* satisfies the "average" user who issues query *q*.

The general problem is NP-hard, but the objective function admits a submodularity structure that can be exploited for the implementation of a good approximation algorithm. Without going into details, the overall idea is to probabilistically classify queries and documents into taxonomy of intents and then use the analogy of marginal utility to determine whether to include more results for an already covered intent. The classification is done by analyzing the log of documents clicked by different users asking queries on the search engine [6]. Again, the quality of results improve as more and more satisfied users ask even more queries and click on results.

3. BEYOND WEB SEARCH

Having described illustrative applications where data externality is already playing a role, we next give some speculative applications of data externality.

3.1 Health

An emergent social phenomenon in healthcare is that individuals are starting to take charge of their own health and trying to avoid needing care in the first place [7]. The tools they use include everything from pedometers/accelerometers that monitor footsteps and motion, to sleep monitors, pulse and heart monitors, and glucose monitors. New systems have started to emerge that will allow people to record everything pertaining to their life and to store all these data in a digital archive, possibly in the cloud [8].

This unprecedented creation of user data can lead to large positive externalities. A simple example is the ability to create demographic specific height and weight charts for infants.[1] Similarly, it has been now determined that the optimum

[1] When our daughter was growing up, her (height, weight) point always fell way below the chart. It was of very little comfort to me or my wife when we were told not to worry since the charts were developed for Caucasians.

cholesterol level for Asian Indians is 150 mg/dL (much lower than 200 mg/dL for Westerners) due to elevated levels of lipoprotein [9]. Such data-driven medical discoveries become feasible with the availability of archives of peoples' digital diaries. Of course, applications will have to be endowed with the right privacy and security capabilities for data sharing for such externalities to be realized.

3.2 Education

Quality of educational material plays a critical role in knowledge acquisition on part of students [10]. Consider an online learning setting in which the students are expected to take a multiple-choice quiz to self-test their understanding of the section. The scores for every question for every student are recorded. This data can now be potentially used to identify those sections of the textbook that might be confusing, and hence merit rewriting.

Postulate: Test score = f (student ability, clarity of material). We have record of test scores for a large number of students. Techniques such as those described in [11] can now be applied to estimate the clarity of the material. For another research effort directed at improving the quality of textbooks through data mining, see [12].

4. RESEARCH OPPORTUNITIES

We next sketch some research opportunities related to data externality.

Architecting for data externality. It is in the nature of externalities that they have tipping points where there is general acceptance and near-universal usage. Search engines are already seeing daily data rates in upwards of tens of terra bytes and the data rates are expected to continue to accelerate. Developing architectures for managing and handling such large data sets falls well within the purview of data researchers.

Privacy, security, confidentiality, trust. Data externality requires technology support for responsible data custodianship, without which the system will come unglued [13]. While the progress on this crucial topic will require further exploration of many streams of ongoing research, it may also provide opportunity for new approaches. For instance, an economic approach to privacy was hinted in [14, Section 7.2.4], which might be worth pursuing further.

Data mining and learning at scale. There is need for exploration along couple of complementary dimensions: How to scale the current solutions to work on qualitatively larger datasets? Do some techniques that were not competitive in past become competitive with the availability of much more data? Do we need to develop altogether new techniques?

Design principles. How does one identify applications that could be amenable to data externality? Are there general principles one must follow for taking advantage of data externality? How to ensure incentive compatibility between those providing data and those benefitting from data spillover?

Intrinsic value of data. The primary way the information services based on data externality become economically viable is from the value extracted by analyzing the spilled over data. It behooves then to think through how one can quantify the value of a collection of data or when can we say that one collection is more valuable than the other. This is terra incognita.

5. CONCLUSIONS

Information services developed to consciously take advantage of data externality is a relatively recent phenomenon. By definition, these services are data-intensive in nature. Most of them have been developed as one-ofs. Finding common abstractions, developing general algorithms, and architecting systems for supporting such services could constitute an exciting research agenda for data researchers.

6. ACKNOWLEDGMENTS

This paper is a synthesis of ideas and discussions in Search Labs.

7. REFERENCES

[1] Laffont, J. J. 2008. Externalities. *In The New Palgrave Dictionary of Economics*, S. N. Durlauf and L. E. Blume, Ed. Second Edition.

[2] Burges, C., Shaked, T., Renshaw, E., Deeds, M., Hamilton, N., and Hullender, G. 2005. Learning to Rank Using Gradient Descent. In *ICML*.

[3] Banko, M. and Brill, E. 2001. Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *ACL*.

[4] Agrawal, R., Halverson, A., Kenthapadi, K., Mishra, N., and Tsaparas, P. 2009. Generating Labels from Clicks. In *WSDM*.

[5] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. 2009. Diversifying Search Results. In *WSDM*.

[6] Fuxman, A., Tsaparas, P., Achan, K., and Agrawal, R. 2008. Using the Wisdom of the Crowds for Keyword Generation. In *WWW*.

[7] Dyson, E. 2010. Heal Thyself. *Project Syndicate* (April 2010). DOI= http://www.project-syndicate.org/commentary/dyson19/English.

[8] Bell, G. and Gemmell, J. 2007. A Digital Life. *Scientific American* (March 2007).

[9] Enas et al. 2001. Coronary Artery Disease in Asian Indians. *Internet J. Cardiology.*

[10] Heyneman, S. P., Farrell, J. P., and Sepulveda-Stuardo, M. A. 1978. Textbooks and Achievement: What We Know. World Bank. PUB HG 3881.5 .W57 W67 No. 298.

[11] Baker, F. B., & Kim, S.-H. 2004. *Item Response Theory: Parameter Estimation Techniques* (2nd ed.). Marcel Dekker.

[12] Agrawal, R., Gollapudi, S., Kenthapadi, K., Srivastava, N., and Velu, R. 2010. Enriching Textbooks Through Data Mining. In *ACM DEV*.

[13] Federal Trade Commission 2010. Protecting Consumer Privacy in an Era of Rapid Change. December 2010.

[14] Agrawal, R., Freytag, J. C., and Ramakrishnan, R. (Eds.) 2004. Data Mining: The Next Generation. Dagstuhl Seminar 04292. July 2004.