Virgil Pavlu December 1st,2010

Research Statement

VISION

A new approach to knowledge. In the last few decades we have witnessed a major shift to a digital world that not only affects all major dimensions of modern civilization -culture, commerce, military and science-, but it completely changes long established norms. Consider the concept of a "library": since ancient times, many organizations collected and archived records, writings and facts into voluminous physical spaces. Today, all the information anyone needs can be stored or accessed on a pocket-size device; for example, all Wikipedia fits into a cell phone while the entire Library of Congress can be stored on a workstation. This is made possible by of major advances in systems, architectures and miniaturization; however we need new tools to make use of the vast majority of data we now have access to. Two newer disciplines are quickly becoming foundations of "modern libraries": Machine Learning [ML] is responsible for mining and creating knowledge from data, and Information Retrieval [IR] is responsible for accessing the data. My main research interests lie in these two fields and in their integration for various problems; I consistently use ML, Information Theory and statistics as a base to approach IR problems, with notable success on those concerning evaluation.

Search engines are much of IR, but much more than meets the naive eye. While the basic concept is search, for it to work, a significant infrastructure must be in place, often more important than the search mechanism: caching and updating very large datasets, making sense of implicit data structure, dealing with billions of queries a day, personalizing the results, etc. My work is focused on search quality, applied Machine Learning, and modeling of text; I have developed algorithms and models for efficient evaluation, estimation of query difficulty, metasearch, IR scoring functions, and exploration of relevant patterns. I also participated to collection-building efforts for research purposes. My thesis was awarded the Northeastern University Dissertation Completion Fellowship for Spring 2008.

Today, search and mining techniques are embedded into all aspects of digital world: Internet search, desktop search "as you type", local network search, search on cell phones, etc.; therefore the importance of their efficacy and efficiency cannot be overstated.

Foundations, theory and practice. My training is that of a theoretician with a strong background in algorithms and mathematics, and with good programming skills; hence my research relies on mathematical principles or derivations, and the implementations are self-coded.

In the Science of Computers, there are many things that work well, and it is my firm belief that there are scientific, rigorous explanations for all of them. I accept that sometimes the explanation could be beyond our reach; nevertheless, a solution that works, but that we don't understand, is just a heuristic for which we should try our best to reason. I want to solve *real* problems; while I know complexity is necessary, I believe in "making everything as simple as possible, but *not* any simpler"(Einstein). I favor seeing before deducing (that is, a good representation may actually show the solution of the problem), hands-on approaches, justifiable outcomes and I consider 'why' and 'how' equally important questions.

PAST AND CURRENT WORK

Does the search engine answer satisfy the information need? As it turns out, this is a difficult question: first, it is hard to encapsulate the questioner satisfaction with the answer in a formula; second, even if we take a certain performance-measurement formula for granted, it would require a enormous human effort to decide if all relevant pieces were retrieved and also which of the pieces retrieved are in fact not relevant. My PhD thesis proposes a new approach to large scale performance measurement, specifically on how to massively reduce this human effort [PHDTHESIS]. Below are brief summaries of several of my past and current projects.

Maximum entropy method. Over the years, many measurements have been proposed for IR performance, each with different requirements, purpose and features. Some of them became increasingly popular, achieving an unofficial status of “gold standard”, but this happened mostly based on intuition rather than formal reasoning. We used the maximum entropy framework to prove that the ”gold standard” measures are indeed the most informative, in a information-theoretic sense [MAXENT].

Geometric representation. For a long time, some of the performance measurements were known to be well correlated, due to empirical evidence. We found a geometric representation of these measures that explained the correlation [RPREC].

Statistical survey applications. It is amazing (naively speaking) how accurately polls can estimate the number of votes a candidate will take across the country, using only about two thousands inquiries; even more, they can give confidence in the predictions. That is because of the mathematical powerhouse of Statistics, but also because of the following design: Say we want to know how many popular votes Sarah Palin would take as candidate to U.S. presidency (note, however, that in U.S. the presidential election is not decided by direct popular vote). It makes statistical sense to question four times more Texans than New Yorkers, that is to take into account the “prior” belief that Sarah Palin is four times more popular in Texas that she is in New York. Independent of this prior holding or not in reality, Statistics weights this four to one ratio against the population total to make sure the estimates are correct (more precisely: unbiased). What the prior is giving us is that, as long as Mrs Palin is more popular in Texas than she is in New York, the estimates will have lower variance than they would if we were to use a uniform survey over all states. This work was funded by NSF 2006-2009.

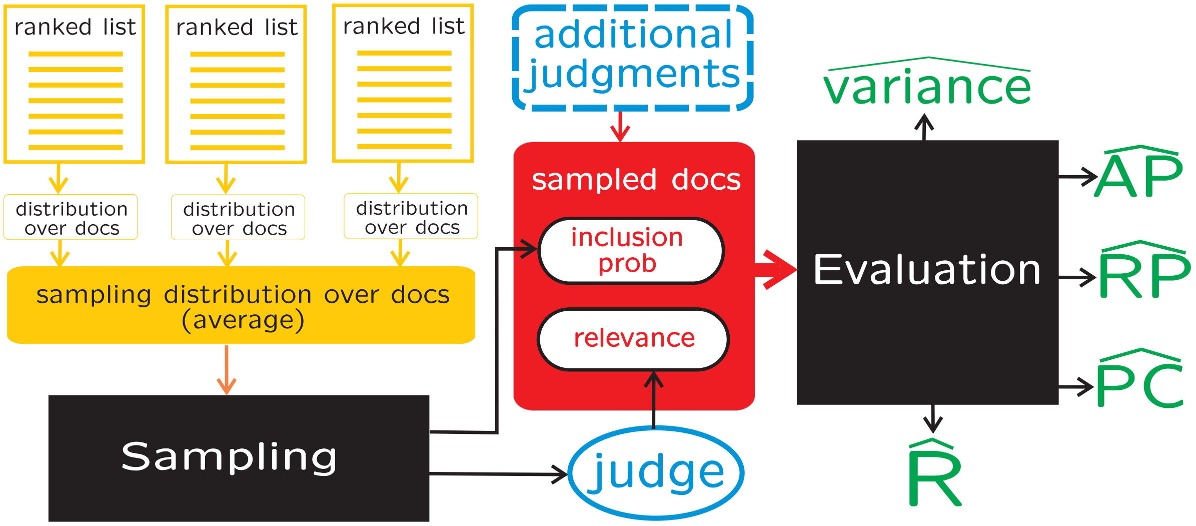

Figure: Survey theory applied to SAMPLING, JUDGING, and EVALUATION. It estimates Average Precision(AP), RPrecision(RP), Precision at cutoff(PC), number of relevant documents(R) and also gives the variance of the estimates.

Similarly, in IR, we can construct a prior of relevance over retrieved documents, and so we can use surveys to estimate many of the popular measures using only 5% of the human effort typically employed [SAMPLING].

TREC, Million Query Tracks 2007-2009. National Institute for Standards and Technology [NIST] sponsors annually the Text REtrieval Conference [TREC], where various research groups run their search engines on given data-collections and queries and later obtain evaluations of performance. This is a critical event, because most of IR research is verified using TREC data. In the closest track to real live search, the "ad-hoc" track, evaluations of some 50-to-100 picked queries consists in human-assessing of about 80,000 documents for relevance, which takes more than 2 man-years effort (which is why only governmental institutions and big corporations can afford such tasks). In 2007, TREC used our survey-based statistical technique (together with an alternative strategy developed at UMass Amherst) on Million Query Track, in order to evaluate an unprecedented number of 1,800 queries . In 2008 and 2009 we kept the basics of the track, and added new tasks (like query predictions) and new analysis (like reusability) [MQ2007-2009]

IR data reusability study. A big question to any major effort of collecting data is whether the data collected can be used more broadly: for example TREC tracks primarily collect documents relevance judgments for evaluating each track submitted IR systems, but many researchers would like to use the data collected in the following years for various other purposes. We simulated an evaluation setup to answer how useful is the collected data in measuring the performance of "brand new" IR systems. [REUSE2010]

Learning to Rank In the past few years, a lot of research has been focused on training search engines (i.e. apply Machine Learning techniques in order to obtain a ranking model, which in turn is used to rank retrieved results for any user query). But how do we train such ranking models? The answer is to use documents marked for relevance by organizations such as TREC in previous years. We looked at the effect various pooling methods (used by TREC) have on the ranking model trained on selected documents, and found that some pooling methods are better than others [LTR2009]. This work is currently being funded by NSF.

Score Distribution Models. Inferring the score distribution of relevant and non-relevant documents is an essential task for many IR applications (e.g. information filtering, recall-oriented IR, meta-search, distributed IR). Modeling score distributions in an accurate manner is the basis of any inference. Thus, numerous score distribution models have been proposed in the literature. Most of the models were proposed on the basis of empirical evidence and goodness-of-fit. In this work, we model score distributions in a rather different, systematic manner. We start with a basic assumption on the distribution of terms in a document. Following the transformations applied on term frequencies by two basic ranking functions, BM25 and Language Models, we derive the distribution of the produced scores for all documents. Then we focus on the relevant documents. In particular, assuming a Gamma distribution for all retrieved documents, we show that the derived distribution for the relevant documents resembles a Gaussian distribution with a heavy right tail [SD2010]

Performance model by class. There are many schemas proposed for search engines, although most of them are variations of no more than about ten fundamentally different classes of approaches. I have early evidence that the performance of certain schemas can be modeled by the class they belong to.

Metasearch using online allocation. Metasearch is the problem of combining the output of several search engines (on the same query); practically, it is an internal mechanism used by all major engines, because many search features and techniques are always combined into a single final output. The classical ML online allocation problem and the Hedge algorithm (also used to combine episodic expert advice) are analogous and therefore easily applicable to the IR problem of metasearch. We obtained, besides metasearch, a way to find very relevant documents and to incorporate feedback into the search strategy, and also a fast method for differentiating the search systems in terms of performance [HEDGE].

Query difficulty estimation is addressing the large diversity of the queries that search engines encounter. Some of them are considerably more difficult than others for various reasons: ambiguity, generality, lack of relevant answers, language constructs confusion etc. From the perspective of the search engine, it would be great if queries could be classified before sending out the answer; this could allow hard queries to have special treatment. We developed a technique based on the information-theoretical Jensen-Shannon Divergence that, given two or more search strategies, estimates query difficulty [QDIFFICULT]. Since usually several searches are executed internally, this is an easy add-on to the overall search procedure.

Log analysis tools can be very useful for computer forensics, assuming an intrusion is detected, especially if the intrusion is of the bad kind. In a scenario where there are millions of records and a few system administrators working around the clock to bring the infrastructure back to a functional and safe state, the ability to spot an anomaly is critical. We proposed an approach based on information theory, together with a log visualization utility; currently work is being done on a plugin for the popular network tool Ethereal [LOGTREE].

FUTURE WORK

My short term goals are focused on several ongoing projects I am involved in; I have some plans for medium term work, and also several long term ideas. Here is a summary of my current and future projects, ordered from short term to long term.

A content-based representation of document relevance A current problem with IR research is that document relevance is understood as a small set of binary (0/1) marked documents, which completely misses any notion of what "relevance" means for that particular user or query. Instead we are working on a model that does not marks documents relevant or not, but instead uses a set of "nuggets" (fragments of documents, sentences, snippets etc) to characterize relevance. Such a model would solve quite easily problems like (1) recall: where a lot of documents are relevant, but relevance there can be inferred from few facts; and (2) reusability: given a collection of nuggets, we can quickly measure performance of any result list, even if the documents retrieved are previously unseen.

Making sense of political bias in news articles We are building a model for predicting political bias of news articles. The project consists of the following modules: (a) a crawler focused on political news, blogs, articles, websites; (b) a NLP-based framework for extracting political predicates in structured form, for example (actor, political issue, quantitative vector) = (GOP leadership, taxes, decrease); and (c) a Nearest-Neighbor mechanism for learning and predicting.

User studies. In order to understand better what users are looking for when using a search engine, we are planing two user studies: (1) a study for identifying relevant information in documents, and verifying that other documents which match this information are valuable; (2) a study designed to measure accurately and realistically measure search engine performance from a user point of view: a measure of cost (time spend reading/browsing) vs utility (how much information is accumulated).

Automatic diversification using named entity recognition Diversity in Information Retrieval is the notion that while a user wants to see relevant results, he prefers those that are somewhat different to the ones already seen. We are working on an automatic process to diversity search engine results using a named entity tagger and an information theoretic framework for tags distribution across documents.

Search engine optimization. An IR performance (or quality) measure can be a key component of a search engine, if internally used as an objective function. Direct learning approaches to search and ranking have been proposed by characterizing this fact [APSVM]. Our study of performance measures combined with our internal-metasearch expertise could potentially lead to a good contribution in this area.

Summarization of search would be of practical interest for a user who just received on his terminal 300,000 results as response for a query. Obviously, such a number of documents is beyond his capability of examination, but say he is willing to spend some time on the results. Can the 300,000 documents be meaningfully summarized into several pages? To make things clear, I am not referring to summarizing each document (this is a well established subfield of IR), but rather summarizing the content of all documents as a whole. My intention is to use survey theory, clustering, and information extraction techniques to achieve the summarization.

New performance measurements. While many search engines present the output as a ranked list, there are approaches based on clustering that usually work like portals: they let the user navigates "a tree" from a given top, and that narrows the area clustered with each click. What would be a good performance measure for this form of output? Minimum Description Length principle could be a good start for solving this problem.

Integration. More access to information is definitely good; but it also produces more chaos, and we humans are excessively good at generating chaos. If all the information will be *consistently* organized, then perhaps the established field of Databases will serve us well enough. Not only this is not the case, but the more we collect data -and we collect at very high rates-, the closer we bring the traditional database era to an end.

What we have is vast amounts of unorganized data: text, audio, video, personal records, fingerprints, datasets and, to make matters worse, proprietary formats. We need ML and IR to manage information in natural form, and we need to integrate them with Databases tools. Mathematics can easily be integrated (because mathematics *is* modeling natural forms), but the SQL language, *as it is,* cannot. Databases need to adapt to the new realities, while ML and IR need to use the existing database infrastructure. This topic is a very exciting area of research, which is getting increasing attention [DBINTEGRATION]

REFERENCES

[PHDTHESIS]

Large Scale IR Evaluation, PhD thesis, August 2008.

[SAMPLING]

* A Practical Sampling Strategy for Efficient Retrieval Evaluation, with Javed Aslam, submitted for review.

* A Statistical Method for System Evaluation Using Incomplete Judgments, with Javed Aslam and Emine Yilmaz, SIGIR 2006.

* A Sampling Technique for Efficiently Estimating Measures of Query Retrieval Performance Using Incomplete Judgments, with Javed Aslam and Emine Yilmaz, ICML 2005 Workshop on Learning with Partially Classified Training Data.

[LTR2009]

Document Selection Methodologies for Efficient and Effective Learning-to-Rank, with Evangelos Kanoulas, Javed Aslam, Stefan Savev and Emine Yilmaz, SIGIR 2009

[SD2010]

Score Distribution Models: Assumptions, Intuition, and Robustness to Score Manipulation, with Evangelos Kanoulas, Javed Aslam, Keshi Dai, SIGIR 2010

Modeling the Score Distributions of Relevant and Non-relevant Documents, with Evangelos Kanoulas, Javed Aslam, Keshi Dai, ICTIR 2009

Variational Bayes for Modeling Score Distributions, with Evangelos Kanoulas, Javed Aslam, Keshi Dai, Journal of Information Retrieval, 2009.

[REUSE2010]

Reusable Test Collections Through Experimental Design, with Ben Carterette, Evangelos Kanoulas, Hui Fang, SIGIR 2010.

[MQ2007-2009]

* James Allan, Ben Carterette, Javed A. Aslam, Virgil Pavlu, Blagovest Dachev, Evangelos Kanoulas: Million Query Track 2007,2008,2009 Overview papers
* Evaluation Over Thousands of Queries with James Allan, Ben Carterette, Javed A. Aslam, Evangelos Kanoulas, SIGIR 2008
* If I Had a Million Queries, with James Allan, Ben Carterette, Javed A. Aslam, Evangelos Kanoulas, ECIR 2009

[QDIFFICULT]

Query Hardness Estimation Using Jensen-Shannon Divergence Among Multiple Scoring Functions, with Javed Aslam, ECIR 2007.

[HEDGE]

A Unified Model for Metasearch and the Efficient Evaluation of Retrieval Systems via the Hedge Algorithm, with Javed Aslam and Robert Savell, SIGIR 2003.

The Hedge Algorithm for Metasearch at TREC 2006, with Javed Aslam and Carlos Rei, TREC 2006.

A Unified Model for Metasearch, Pooling, and System Evaluation, with Javed Aslam and Robert Savell, CIKM 2003.

Measure-based Metasearch,with Javed Aslam and Emine Yilmaz, SIGIR 2005.

[LOGTREE]

Semi-supervised Data Organization for Interactive Anomaly Analysis, with Javed Aslam and Sergey Bratus, ICMLA 2006.

[MAXENT]

The Maximum Entropy Method for Analyzing Retrieval Measures, with Javed Aslam and Emine Yilmaz, SIGIR 2005.

[RPREC]

A Geometric Interpretation of R-precision and Its Correlation with Average Precision, with Javed Aslam and Emine Yilmaz, SIGIR 2005.

[APSVM]

Yisong Yue, Thomas Finley, Filip Radlinski, Thorsten Joachims: A support vector method for optimizing average precision, SIGIR 07, doi = {http://doi.acm.org/10.1145/1277741.1277790}

Thorsten Joachims: A support vector method for multivariate performance measures, ICML 05 doi = {http://doi.acm.org/10.1145/1102351.1102399}

[DBINTEGRATION]

W. Bruce Croft and Hans-J. Schek: Introduction to the special issue on database and information retrieval integration, http://www.springerlink.com/content/x7082416710h5357/