Revisiting the Foundations of IR: Timeless, Yet Timely

Paul B. Kantor
Rutgers University and CCICADA
96 Frelinghuysen Dr., Piscataway NJ 08854-8018 USA
732-322-8412 paul.kantor@rutgers.edu

ABSTRACT

As we face an explosion of potential new applications for the fundamental concepts and technologies of information retrieval, ranging from ad ranking to social media, from collaborative recommending to question answering systems, many researchers are spending unnecessary time reinventing ideas and relationships that are buried in the prehistory of information retrieval (which, for many researchers, means anything published before they entered graduate school).

Much of today's received wisdom may be nothing more than the fossilized residue of lively debates concerning such things as . estimation of value and evaluation of systems. Returning to those discussions may open the door to genuinely new insights.

On the other hand, of the ideas that surface as "new" in today's super-heated research environment have very firm roots in earlier developments in fields as diverse as citation analysis, statistics, and pattern recognition. The purpose of this tutorial is to survey those roots, and their relation to the contemporary fruits on the tree of information retrieval, and to separate, as much as is possible in an era of increasing commercial secrecy about methods, the problems to be solved, the algorithms for solving them, and the heuristics that are the bread and butter of a working operation.

Among the important new topics whose foundations will be explored are the use of social media in search and advertising, and the growing management of personal image collections for search and for commercial purposes.

While some might think that an examination of the roots is of merely historical interest, it has practical value as well. When you know which earlier research has provided the origins for the things that you are interested in, you can use that fact to trace its other descendents, and often find rich and rewarding ideas in a literature that you would not normally reach, because it was not considered important by your instructors when you were learning about the problems. In addition to pattern recognition and citation analysis, the tutorial will also expose and review some of the relations to the fields of statistics and operations research.

Participants will become familiar with roots in Pattern Analysis, Statistics, Information Science and other sources of key ideas that reappear in the current development of Information Retrieval as it applies to Search Engines, Social Media, and Collaborative Systems. They will be able to separate problems from algorithms, and algorithms from heuristics, in the application of these ideas to their own research and/or development activities. Course materials will be made available on a Web site two weeks prior to the tutorial. They will include links to relevant software; links to publications that will be discussed; and mechanisms for chat among the tutorial participants, before, during and after the tutorial.

Specific Topics will include:

1. The earliest "automated retrieval;

2. Vectors and Logarithms: their pragmatic origins;

3. Probabilistic approaches. A frequentist foundation;

4. The quest for a theoretical foundation;

5. Generative approaches;

6. Network approaches;

7. Binding approaches together.

Certain materials that are very hard to find, including foundational IBM technical reports by H.P. Luhn and I.J. Good will be scanned and placed on a web site for the tutorial. Since this will not be a public site, we will need some information on advance registration to add participants to the site.

Categories and Subject Descriptors
H.3.0 General; H.3.3 Information Search and Retrieval

General Terms: Algorithms, Measurement, Performance, Theory

Keywords: Foundations of Information Retrieval

1 Overview and Rationale

The purpose of this tutorial is to survey those roots, and their relation to the contemporary fruits on the tree of information retrieval, and to separate, as much as is possible in an era of increasing secrecy about methods, the problems to be solved, the algorithms for solving them, and the heuristics that are the bread and butter of a working operation.

2 Relevance to the Information Retrieval Community

While some might think that an examination of the roots is of merely historical interest, it has practical value as well. When you know which earlier research has provided the origins for the things that you are interested in, you can use that fact to trace its other descendants, and often find rich and rewarding ideas in a literature that you would not normally reach, because it was not considered important by your instructors when you were learning about the problems. In addition to pattern recognition and citation analysis, the tutorial will also expose and review some of the relations to the fields of statistics and operations research.

SIGIR '15, August 09-13, 2015, Santiago, Chile
ACM 978-1-4503-3621-5/15/08. http://dx.doi.org/10.1145/2766462.2767878

3 Course Objectives

Participants will become familiar with roots in Pattern Analysis, Statistics, Information Science and other sources of key ideas that reappear in the current development of Information Retrieval as it applies to Search Engines, Social Media, and Collaborative Systems. They will be able to separate problems from algorithms, and algorithms from heuristics, in the application of these ideas to their own research and/or development activities.

4 Outline of the Tutorial

1. The earliest "automated retrieval:
    - A. Origins in the American Chemical Society research
    - B. Kent experiments;
    - C. Sets and modifications
        - (i) Boole combinations
        - (ii) Quorum type rules rules;
    - D. Key word selection
    - E. *Luhn; ranking texts*
        - (i) the origins of *tf*
        - (ii) the origins of *idf*
    - F. Ranking and retrieval
2. Vectors and Logarithms: their pragmatic origins
    - A. Weighting the Importance of Features
    - B. Controlling for document size
        - (i) Salton and the ray space approach
        - (ii) Fox's metrics
        - (iii) "Pivoting and Singhal
    - C. The unreasonable introduction of logarithms
    - D. Is there a "vector space of documents and queries"
        - (i) Queries as dual spaces
        - (ii) Non Euclidean metrics
    - E. Patterns in spaces
        - (i) Linear subspaces
        - (ii) Manifolds
        - (iii) Relative densities – relation to sensor problems
3. Probabilistic approaches. A frequentist foundation
    - A. Maron and Kuhns
    - B. One user among many – the "system orientation?"
    - C. One document and one users – subjective probability
    - D. One user and a feature-based class of documents
        - (i) The "cluster hypothesis"
        - (ii) The "smoothness of relevance"
    - E. Research in Pattern Recognition
        - (i) Estimating densities empirically
4. The quest for a theoretical foundation:
    - A. I.J. Good and the weight of evidence. The *"bin"*
        - (i) odds of relevance
        - (ii) the relevance of log-odds
    - B. Robertson and Sparck-Jones
        - (i) explaining" term weighting
        - (ii) estimating parameters from data
    - C. Naïve but successful
        - (i) Is there a "deep reason"?
        - (ii) Robertson and "BM25".
        - (iii) Hierarchical models
5. Generative approaches: Language models and topic modeling
    - A. A new layer of formalism
    - **B.** hidden relations to the older views
    - C. characterizing a literature, a topic or an interest
    - D. Conjugate distributions and hyperparameters
    - E. Mixtures of distributions
    - F. Random numbers of random variables.
6. Network approaches:
    - A. Graph theory and IR
    - A. Origins in citation studies
        - (i) Impact Factor
        - (ii) Network instabilities
        - (iii) Co-citation analysis
    - B. Warren; Garfield; Small; Narin;
    - C. Kleinberg. Directed graphs. Hubs and authorities
    - D. Brin & Page: Page Rank
    - E. Ask.com and the "media war"
    - *F.* Networks social and other for information finding
        - (i) Folksonomies and Tagging
        - (ii) Association of persons
        - (iii) Privacy and social stability.
7. Binding approaches together:
    - A. No one method works
    - B. The kinds of synthesis
        - (i) Feature expansion
        - (ii) Rule combination
    - C. The scope of synthesis
        - (i) Global
        - (ii) User-dependent
        - (iii) Task-dependent
    - D. Usage as a meta-feature
        - (i) Covert collection
        - (ii) Overt cooperation
8. Wild Speculations Can there be a theory of how information and users interact? Brains without brain models?
    - A. Quantum formulations: information collapses the users' wave function
    - B. HOTT (HOmotopy Type Theory): user experience as a path in info. space
    - C. An Appreciation of IJ Good.

5 Course Materials

Course materials will be made available on a Web site two weeks prior to the tutorial. They will include links to **relevant software**; links to **publications** that will be discussed; and mechanisms for **chat** among the tutorial participants, before, during and after the tutorial.

6 Some Relevant Monographs and Papers

[1] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern classification. 2001.

[2] Geman, S. & Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (6), pp.721–741.

[3] I.J. Good. Probability and the Weighing of Evidence. Griffin, 1950.

[4] I.J. Good. Rational decisions. Journal of the Royal Statistical Society. Series B (Methodological), 14(1):107{114, 1952.

[5] I.J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. The Annals of Mathematical Statistics, 34(3):911{934, 1963.

[6] IJ Good and RA Gaskins. Nonparametric roughness penalties for probability densities. Biometrika, 58(2):255{277, 1971.

[7] T. Hastie, R. Tibshirani, and J.H. Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer Verlag, 2009.

[8] H.P. Luhn. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4):309{317, 1957.

[9] H.P. Luhn. A business intelligence system. IBM Journal of Research and Development, 2(4):314{319, 1958.

[10] H.P. Luhn. The automatic creation of literature abstracts. IBM Journal of research and development, 2(2):159{165, 1958.

[11] P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. Ann. Probab. Volume 18, Number 3 (1990), 1269-1283

[12] Bishop, Christopher M. Pattern Recognition and Machine Learning. 1st ed. 2006. Corr. 2nd printing, 2006, XX, 740 p. 304 illus. in color. Springer

7 Selected Readings

Certain materials that are very hard to find, such as some papers by H.P. Luhn and I.J. Good will be scanned and placed on a web site for the tutorial. Since this will not be a public site, we will need some information on advance registration to add participants to the site.

ACKNOWLEDGMENTS

Thanks to John Prager and the Library at IBM for providing the classic papers.