ACM SIGIR 2000
Workshop on Mathematical/Formal Methods in Information Retrieval (MF/IR 2000)
July 28, 2000, Athens, Greece

Chairs: Sandor Dominich, Mounia Lalmas, Keith van Rijsbergen

> "The development of Western Science relies on the invention of formal logical system (Euclidean Geometry by Greek philosophers), and on the discovery to find out causal relationships by experimentation. " (Albert Einstein)

Introduction

The purpose of the ACM SIGIR 2000 Workshop on Mathematical/Formal Methods in Information Retrieval (MF/IR 2000) was to promote discussion and interaction among researchers primarily concerned with mathematical (formal) aspects of IR, especially in the areas of foundations, models, properties, and structures, both from a theoretical and practical point of view.

In this respect the workshop was very successful.

As this SIGIR Conference showed too, there is a vast amount of experimental and theoretical knowledge accumulated in IR so far, and the dynamics with which new results are obtained and new knowledge is gained is impressive, especially with the advent of the Internet and World Wide Web (WWW).

Any IR system is trying to answer — or aims at answering — a user's request by way of retrieval, which is implemented using algorithms based on some mathematical model of the main entities involved (objects to be searched, user's request, relevance to query, and so on). The mathematical methods used so far in retrieval modelling rely on vector space theory, probability theory, classical set theory, Boolean logic, fuzzy set theory, topology, algebra, matroid theory, recursion theory, rough sets, decision theory. Beside the mathematical methods, different alternative ideas, methods and techniques (e.g. knowledge base, artificial neural networks, etc.) help improve retrieval.

An important debate topic during the day was whether the mathematical results achieved so far in IR could be organised into a coherent theoretical framework, and what new knowledge could mathematics bring to IR. Also, an important question which emerged during the workshop was whether mathematical/formal research can stand as a specialised research area or discipline of IR.

The workshop was very successful in this respect, too, because answers were found to these questions — they are presented in the Conclusions.

General

The ACM SIGIR 2000 MF/IR Workshop was held on the 28th of July in Athens (Greece) within the ACM SIGIR 2000 Conference, and included four sessions as follows:

1. Algorithms, Methods,
2. Concepts, Vocabularies, Thesauri,
3. Profiling, Human Aspects,
4. Query Expansion.

There were ten speakers in all, who were selected based on a peer review process preceding the workshop. Their papers are published in a special issue of Technology Letters (volume 4, number 1, 2000; ISSN 1369-3735), the official scientific journal of Buckinghamshire Chilterns University College (United Kingdom).

There were 24 people in the audience from around the world, and this number was sometimes higher as a result of the oscillation of inter-workshop attendance.

The atmosphere of the workshop was at a high scientific level, and allowed for the exchange of many interesting ideas, and intense debate especially on the role mathematics plays in IR.

Sessions

In what follows summaries of presentations are given.

1. Algorithms, Methods,

Why the Vector Space Model Works? P. Bollmann-Sdorra, T. Graepel, R. Herbrich, and V. Raghavan.

Bollmann-Sdorra argued that the vector space model of IR was based on a sound mathematical theory: it had an axiomatic foundation, several similarity measures could be transformed into a linear form, and there was a probabilistic interpretation of this model. Thus he proved that the vector space model had its own well-defined mathematical formalism and theory.

Formal Foundation of Classical Information Retrieval. S. Dominich.

Dominich showed that the two classical models of IR (vector space, probabilistic) had a unified mathematical definition from which all known mathematical results and properties could be formally derived. He also gave a formal definition for the Boolean model, and showed that this model was not a distinct elementary model of IR. Thus IR

could become a formal (mathematical) discipline, too (of course besides belonging to information and social sciences). In IR practice, however, it is more convenient to use and see these models as distinct models.

Efficient Algorithms for Ranking Documents Represented As DNF Formulas. D. E. Losada, and A. Barreiro.

Losada described algorithms with which the measure of the uncertainty in a logical model could be evaluated in polynomial time (as opposed to previous exponential times). His idea was to represent both documents and queries in a DNF of terms.

2. Concepts, Vocabularies, Thesauri

Partial Boolean Algebras As Models for Thesaurus Integration. C. Ferigato.

Ferigato offered a view on IR based on abstract algebra. He presented his theoretical research on using Boolean Algebras to model thesauri, as well as different operations such as indexing, integration, retrieval, browsing.

Adaptive Concept-based Retrieval Using Neural Network. M. Kim, and V. Raghavan.

Kim suggested a multi-layered perceptron neural network in which a back-propagation learning rule (relevance feedback) was used to adjust the weights of an AND/OR tree which represented query concepts (in the RUBRIC system). He reported on experiments which showed higher performance (in terms of precision and recall) than in the original RUBRIC approach.

Rough Sets for Mining Vocabulary for Information Retrieval. P. Srinivasan, and D. H. Kraft.

Srinivasan was concerned with applying the theory of rough sets to IR. She presented the theoretical foundations of how a combination of concepts and techniques based on fuzzy sets and rough sets can be applied to IR. Such a combined framework allows for document and query representations and modifications (via vocabulary mining) in such a way that system performance can be enhanced. She also reported on experiments carried out using Medline.

3. Profiling, Human Aspects

Texture, Human Perception, and Information Retrieval Measures. J. S. Payne, L. Hepplewhite, and T. J. Stonham.

Payne concentrated on performance measures in image IR. She reported on experiments carried out in order to calculate, compare and discuss standard performance measures in image retrieval. Brodatz textures were used, and relevance assessments were gathered

from 30 volunteers. Her finding was that the ordering of the retrievals should also affect system usability.

A Bayesian Approach to User Profiling in Information Retrieval. S.K.M. Wong, C.J. Butz.

Butz suggested theoretical foundations for the usage of Bayesian and Markov probabilistic techniques in IR. Relevance feedback was modelled in a Bayesian network which can be refined, and which thus yields a user profile. This could be used in ranking new documents in subsequent retrieval or filtering.

4. Query Expansion

Describing Query Expansion Using Logic-induced Vectors of Performance Measures. M. H. Heine.

Heine focused on how vector theory can be applied to represent effectiveness. He suggested the usage of vectors to model the influence of query expansion (term addition) and choice of logic operator (in Boolean retrieval) on retrieval effectiveness. This influence took the form of changes in direction and magnitude of vectors as well as that of hyper-surfaces. He also reported on experiments carried out using Medline, and on a software he developed to visualise such changes.

CLIR As Query Expansion As Logical Inference. J.-Y. Nie.

Nie shed a new formal view upon cross-language retrieval (CLIR). He suggested a unified theoretical framework. He argued that query translation was a special case of query expansion, which, in turn, could be formulated as a logical inference. He presented the formalism of this new framework, and reported on experiments carried out by implementing this approach.

Conclusions

The topics that raised discussions and lively debate were as follows.

> It became clear that linearity and a preference relationship played a central role in the vector space model. Linearity was important with respect to similarity measures; Jaccard's coefficient could also be transformed into a linear form. The transitivity/intransitivity of the preference relationship (for ranking retrieved documents) is strongly related to the user's relevance judgment.
>
> The unified mathematical definition of the vector space and probabilistic models allowed for deriving (formally) all known mathematical properties and results of these two models, and also of the Boolean model. Thus the basis for a coherent framework was laid down. The connection between the p-norm and this framework was identified as a further research point.

Because learning was unstable at lower weight values in the adaptive concept-based model, experiments with different other versions of back prop were suggested.

A possible connection between IR models based on Boolean Algebras, and those based on rough sets was considered worth being looked at. Both models made use of thesaurus operations, and the model based on Boolean Algebras (a model purely theoretical) could thus become applicable.

The topic of relevance raised long discussions (which is perhaps not surprising, as it is well known that relevance is a crucial modelling parameter in IR), especially as regards retrieval measures using the Brodatz textures.

The possibility to express changes of effectiveness using changes in vectors raised the question of whether tensor theory could be involved with the aim to obtain a more coherent expression of changes in effectiveness.

The workshop helped us better appreciate the role mathematics, and formal methods in general, play in IR. The following results were identified at this workshop as important and novel results brought by mathematics to IR.

The vector space model has its own well-defined axiomatic mathematical theory.

The three classical models of IR (vector space, probabilistic, Boolean) have a unified mathematical foundation, and thus theory.

The underlying formal difference between the vector space and probabilistic models is commutativity (i.e., whether the relevance measure is a commutative function or not).

Because of the huge numbers on the Web (total number of documents stored, number of documents returned) a new way of conceiving retrieval was raised, namely whether retrieval could be conceived as a process of generating a (convergent) sequence of documents which tends towards (has as a limit) the query. Thus the very powerful mathematical concepts of continuity and limit could be involved (they have not been involved yet in IR), and thus the possibility to elaborate new effective models could be investigated.

The formal structure (as a network) of IR models, the relationships between them have become clearer. Although in IR practice many different models are used, and it is convenient to see them as distinct models, formally there are two classes of models: the non-classical models, and the classical models which have one common root, of which the alternative models are enhancements.

During the workshop it became clear that the mathematical results obtained so far in IR can be gathered into a coherent theoretical framework, that mathematics brings new and important knowledge to IR, and that mathematical/formal research in IR can stand as a specialised research area of and as an applied mathematical discipline in IR.

The chairs of this workshop would like to take this opportunity, too, to thank ACM SIGIR, Athens University (Greece), Buckinghamshire University College (United Kingdom), University of Veszprem (Hungary), editorial board of Technology Letters for their help, and to thank all workshop participants, program committee members, reviewers for their work and comments.