Information Retrieval and Applications of Graphical Models (IRGM 2007)

Juan M. Fernández-Luna
Departamento de Ciencias de la Computación e Inteligencia Artificial.
Universidad de Granada, Spain
jmfluna@decsai.ugr.es

Benjamin Piwowarski
Yahoo! Research Latin America
bpiwowar@yahoo-inc.com

Juan F. Huete
Departamento de Ciencias de la Computación e Inteligencia Artificial.
Universidad de Granada, Spain
jhg@decsai.ugr.es

1 Introduction

Probabilistic models constitute an important kind of Information Retrieval (IR) model. They have been long and widely used [2], and offer a principled way of managing the uncertainty that naturally appears in many elements within this field. Nowadays, the dominant approach for managing probability within the field of Artificial Intelligence is based on the use of Bayesian Networks [15, 9], and these have also been used within IR as extensions of classical probabilistic models.

Broadly speaking, a Graphical Model (GM) [10, 17] consists of a qualitative part, a graph, which may be directed and acyclic in the case of Bayesian Networks (BN), and a quantitative one, a collection of numerical parameters. An example of the latter are the conditional probability tables for BNs. The knowledge represented in the graphical component is expressed in terms of (in)dependence relationships between variables. These relationships are encoded by means of the presence or absence of links between nodes in the graph. The knowledge represented in the numerical part quantifies the dependences encoded in the graph, and allows GM to be used as probabilistic models.

Important savings in storage requirements are obtained because independences allow a factorization of the global numerical representation (the joint probability distribution for BNs). They

also allow probabilistic inference (computation of posterior probabilities) to be performed, efficiently in many cases. Their flexibility allow them to encompass many (probabilistic) models in IR and the important development of learning techniques also offers an opportunity to set automatically the models parameters or detect the dependencies between the variables of the model. On the whole, GMs provide a very intuitive graphical tool for representing available knowledge.

Although Bayesian networks were first applied to IR towards the end of the 80s [7, 6], they were more widely used in the next decade following the birth of the Inference Network Model [18]. Since then, many models and applications have been developed, showing that these probabilistic graphical models are suitable to be employed in IR. Recent works include:

- Structured (XML) IR and BN [13, 16, 4].
- Classification of semi-structured documents [5].
- Document reformulation [19].
- Spam detection[12] ...

The interested reader can also found a valuable source of information in the special issue of Information Processing & Management on "Bayesian networks and Information Retrieval" [3].

The development of solutions for IR problems has been very challenging and imaginative, because of two, a priori, drawbacks of BNs, specifically, and of GMs in general: Learning, and to a large extent inference, are time consuming, and are in the general case a NP-hard problem. Therefore, researchers have been looking for solutions to IR problems that take the most of these probabilistic graphical models, but at the same time, have been looking for alternatives to overcome these problems, making them efficient and effective tools.

Meanwhile, in the field of GMs, works have focussed on two main aspects: New inference and learning algorithms have been proposed – allowing more complex models to be used – and new GMs (like e.g. dynamic BN [17]) and new techniques (like the use of Fisher Kernel Discriminants [5] in the context of sequence categorization) have appeared. These developments offer new opportunities for innovative IR applications.

The main goal of this workshop wants to be a common space where researchers can show their innovative GMs applications to the field of IR, in its wide problem space, opening a new discussion forum. Graphical Models include Bayesian Networks, possibilistic networks, Markov networks, dependence graphs, influence diagrams, probability trees, decision trees, and Fisher Kernel Discriminants, among others.

The goal of this workshop is to strengthen up the community, specially young researchers who were encouraged to present their current research, obtaining a valuable feedback from the interested audience. We hope this workshop will also offer a good view of state-of-the-art of this innovative field, presenting interesting solutions to classic and new IR problems. We present in the next section the papers that were accepted for the first edition of this workshop, as well as the keynote speech which opens the workshop.

Paper name	Model/methods	Task
Semantic hashing	Two layer undirected, Restricted Boltzman Machine, Gradient ascent	Finding similar documents
Stacked Dependency Networks for Layout Document Structuring	Relational Dependency Networks	Document structuring and annotation
Using Scatterplots to Improve Naïve Bayes Text Categorization and Retrieval	Naive Bayes Classifier	Text classification
Possibilistic Networks for Information Retrieval	Possibilistic Network	Ad-hoc retrieval
Using Structural Content Information for Learning User Profiles	Bayesian Network	Content-based recommendation

Table 1: Summary of papers

2 Review of accepted papers

Table 1 summarizes the different methods used in the six papers presented in the workshop and the problem for which they have been applied. As one can notice, there is a interesting selection of graphical models, as there is a wide variety of them. A good range of different different open problems found in the Information Retrieval field are addressed. The summaries of all the papers are presented more in details in this section.

The paper entitled "**Semantic Hashing**", by Ruslan Salakhutdinov and Geoffrey Hinton, presents a probabilistic approach to dimensionality reduction of the documents vector space. The authors use an undirected graphical model (a generalization of a Boltzmann machine [8]) whose weights are learned via a two-stage process, whereby the second algorithm fine tunes the results by the first one. The deepest layer of the graphical model is a set of binary variables (128) that can be interpreted as a hashing code for a document. This code has the great advantage of bearing semantic information which used for fast document retrieval: documents similar to a "query document" can be found by considering documents whose code differs only for a few bits from the query. The collection used are categorization datasets: A document is is "relevant" to a query document, if it belongs to the same category. The method is found to be both fast (at retrieval time) and performant (when compared to LSA for instance).

The paper entitled "**Stacked Dependency Networks for Layout Document Structuring**" written by Boris Child-lovskii and Loïc Lecerf, deals with the problem of structuring and annotating document which present an clear internal structure (for example, metadata extraction from electronic and scanned documents). The analysis of the document gives as output a set of elements and its type (element classification), as well as the relationships among them (link classification).

The documents are represented by means of a graph structure. Relational Dependency Networks (RDN) [14] are used to make inference, where the classic inference in this type of graphical

model, Gibbs sampling is substituted by a variant of RDN called Stacked Dependency Networks, due to the slow convergence presented by the former.

The experimentation compares Stacked Dependency Networks with several variants of Gibbs sampling on two different collection. They show that stacking method are much faster in terms of the inference.

The problem of document visualization applied to automatic text classification is explored in "**Using Scatterplots to Improve Naïve Bayes Text Categorization and Retrieval**" by Giogio Maria Di Nunzio. Mainly framed in the context of text categorization, the paper describes a visualization technique by which a document is represented in the Cartesian plane. This is a useful technique to try to understand the relationships among categories of documents as well as to serve the user as a visual tool in order to analyze the performance of the classifier.

The technique proposed in this work is applied to the Naïve Bayes classifier [11], specifically to the well-know Bernoulli and multinomial variants. In order to map a document belonging to a category to a two-dimensional space, the author has to re-write the decision formulation of the classification of both models, obtaining the corresponding two coordinates.

The experiments run on Reuters-21578 and RCV1 supports the idea of representing the probabilities of the documents obtained by these probabilistic classifiers in the Cartesian space, as well as shows that with this technique the behavior of the classifiers in certain conditions could be visually interpreted.

Finally, the re-formulation of the problem in terms of not only classification but retrieval, with the Binary Independence Model, is outlined, introducing an interesting application of the method.

In "**Possibilistic Networks for Information Retrieval**", M. Boughanem, Asma Brini and Didier Dubois, present an IR model based on possibilistic networks [1], characterized by means of a DAG encoding the dependence and independence relationships and a set of conditional possibility matrices encoding the strength of the dependence relationships. The use of Possibility theory allows the model to separate the reasons for rejecting a document as irrelevant (taking into account the possibility values) from the reasons for selecting a document as relevant (by means of the necessity values). This dichotomy is obtained by distinguish between terms which are possibly representative (in general, those terms appearing frequently in a document) and those which are necessarily representative (a term in a document with high discriminant value, i.e. appearing in few documents in the whole collection).

Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete and Miguel Rueda-Morales, in "**Using Structural Content Information for Learning User Profiles**" consider the problem of learning a user profile which captures the mechanisms used for rating an item, in the context of content-based Recommender Systems. In order to define the profile two main assumptions have been considered: On the one hand, a product is described using a well defined set of categories and, on the other hand, the user rating for an item is obtained as a mixture of the ratings for each one of the different structural components. Then, taking into account this assumptions a new user profile is defined, by means of a Bayesian network [15], including two different parts: the first represents how the user should rate each content component, and the second represents how this information might be combined in order to get the global rate. Since the users does not usually expresses their opinions for those structural components, this profile is learnt using the Expectation-Maximization algorithm. The experiments designed to validate the model with a synthetic rate database shows a good behavior and interesting possibilities.

3 Keynote Speech

The invited talk will be given by Yi Zhang, assistant professor at the University of California, Santa Cruz. She obtained its PhD in 2005 at the Carnegie Mellon University, with a dissertation about Graphical Models applied to Information Filtering. The application of these tools to Information Retrieval problems has been the main research line of this young researcher, which presents a very promising research career. From these lines, we would like to thank Yi Zhang publicly for having accepted our invitation to give the opening talk.

The keynote speech is entitled "**Personalized Proactive Information Retrieval with Bayesian Graphical Models**". In this talk, Dr Zhang deals with the problem of alternative information sources, which produce problems in the users, as they do not look up in the correct one when try to satisfy their information needs.

One of the solutions provided to overcome this problem is the use of information retrieval agents, which interacting with the users and continuously observing the different information sources, give valuable information by warning the user about a piece of news in which she/he could be interested. This task comprises typical Information Retrieval tasks as relevance feedback, adaptative or collaborative filtering. She explains how Bayesian decision theory and Graphical Models could be an interesting combination to face the underlying problems of learning, knowledge representation and decision making.

4 Development of the Meeting

The workshop was held at University of Amsterdam during the morning of friday July 27th. There were a total of 15 participants, among speakers, organizers and general attendees.

After an introduction of the workshop by the organizers, the keynote speaker, Yi Zhang offered a very interesting talk, where an example of the application of a Probabilistic Graphical Model was introduced.

Initially, each paper was given a 20 minutes slot, including questions. At the end of the 5 presentations, a period of almost 30 minutes for general discussion was scheduled. But in practice, the discussions took place at the end of each talk (and in some moments during each talk itself) by the questions and comments made by the participants. In this line, participants required explanations, proposed modifications and offered their opinions about the problems, models, solutions and experimentations shown in the papers. The aim of the workshop was to promote the exchange of ideas and impressions, so the chair of the session did not take into account the time assigned to each paper. In general, we think all the authors presenting papers benefit from the workshop.

Finally, at the end of the meeting there was enough time for general discussion. The participants agreed that the papers presented were very interesting and had a high quality. The main drawn conclusions were that in general, Graphical Models offer an interesting framework for modelling and solving a wide range of IR problems, but they usually must be adapted to the specific problem, in order to improve their efficiency.

With respect to organization of a second workshop, all the participants agreed that this sort of research meeting should be always attached to a major event, as it would attract more people. Also the majority feeling was that this event should be biennial.

5 Edition of a Special Issue

During the previous week to the workshop, the organizers were negotiating a special issue of the Journal of Approximate Reasoning (IJAR) with a selection of the best papers coming from extensions of workshop papers. The special issue was accepted in September and all the papers accepted for presentation have been invited to submit extended versions of the original works. Also some papers from the SIGIR conference whose main topic was Graphical Models were invited. The special issue is expected to be in print by the last quarter of 2008.

6 Conclusions

From the submitted papers, we may draw some conclusions for this first edition of the workshop:

- Graphical Models in general can be applied to a wide range of problems in the field of Information Retrieval.
- They show a great representation capacity, deal with problems pervaded with uncertainty, and offer powerful inference mechanisms.
- There is a diverse and active community of researchers working with Graphical Models.
- Although, it is usually said that inference in certain types of Graphical Models, as the case of Bayesian networks is, is a very heavy task, specially in the Information Retrieval field, where there exists a really great information volume, the solutions provided in the papers are examples of how this general drawback can be overcame for each specific problem.

7 Acknowledgments

We would like to thank all the members of the Program Committee their support and advice during the workshop proposal making, and their professionalism in the review phase, both meeting the tight review deadlines, and for the useful comments provided to the authors of submitted papers. The list of PC members is the following:

- Peter Bruza (Queensland University of Technology, Australia).
- Eric Horvitz (Microsoft Research, USA).
- Gianni Amati (Fondazione Ugo Bordoni, Italy).
- Berthier Ribeiro-Neto (Universidade Federal de Minas Gerais, Brazil).
- Iadh Ounis (University of Glasgow, United Kingdom).
- Luis M. de Campos (Universidad de Granada, Spain).
- Didier Dubois (Université Paul Sabatier, France).

- Pável Calado (Instituto Superior Tecnico, Portugal).
- Ludovic Denoyer (LIP6, University Paris 6, France).
- Yi Zhang (University of California, Santa Cruz, USA).
- Alfonso E. Romero (Universidad de Granada, Spain).
- Donald Metzler (University of Massachusetts, USA).
- Dunja Mladenic (Jozef Stefan Institute, Slovenia).

References

[1] D. Benferhat, L. Garcia, and H. Prade. Possibilistic logic bases and possibilistic graphs. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 57–64, 1999.

[2] F. Crestani, M. Lalmas, C. J. Van Rijsbergen, and I. Campbell. "is this document relevant?... probably": A survey of probabilistic models in information retrieval. *ACM Computing Surveys*, 30(4):528–552, Dec. 1998.

[3] L. M. de Campos, J. M. Fernández-Luna, and J. F. Huete. Bayesian networks and information retrieval: an introduction to the special issue. *Information Processing and Management.*, 40(5):727–733, 2004.

[4] L. M. de Campos, J. M. Fernndez-Luna, and J. F. Huete. Using context information in structured document retrieval: an approach based on influence diagrams. *Information Processing and Management*, 40(5):829–847, 2004.

[5] L. Denoyer and P. Gallinari. Bayesian network model for semi-structured document classification. *Information Processing and Management.*, 40(5):807–827, 2004.

[6] M. E. Frisse. Searching for information in a hypertext medical handbook. In *HYPERTEXT '87: Proceeding of the ACM conference on Hypertext*, pages 57–66, New York, NY, USA, 1987. ACM Press.

[7] M. E. Frisse and S. B. Cousins. Information retrieval from hypertext: update on the dynamic medical handbook project. In *HYPERTEXT '89: Proceedings of the second annual ACM conference on Hypertext*, pages 199–212, New York, NY, USA, 1989. ACM Press.

[8] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(5):1711–1800, 2002.

[9] F. V. Jensen. *An introduction to Bayesian Networks.* UCL Press, London, England, 1996.

[10] M. Jordan. Graphical models. *Statistical Science*, 19:140–155, 2004.

[11] D. D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 4–15, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[12] V. Metsis, V. Androutsopoulos, and G. Paliouras. Spam filtering with naive bayes – which naive bayes? In *Third Conference on E-mail and Anti-Spam*, 2006.

[13] S. H. Myaeng, D.-H. Jang, M.-S. Kim, and Z.-C. Zhoo. A flexible model for retrieval of SGML documents. In W. B. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 138–140, Melbourne, Australia, Aug. 1998. ACM Press, New York.

[14] J. Neville and D. Jensen. Dependency networks for relational data. In *Proceedings of the IEEE Data Mining*, 2007.

[15] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[16] B. Piwowarski and P. Gallinari. A bayesian network for XML information retrieval: Searching and learning with the INEX collection. *Information Retrieval*, 8(4):655–681, December 2005.

[17] P. Smyth. Belief networks, hidden markov models, and markov random fields: a unifying view. Pattern Recognition Letter, 1998.

[18] H. R. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions On Information Systems*, 9(3):187–222, 1991.

[19] X. Wei and W. B. Croft. Modeling term associations for ad-hoc retrieval performance within language modeling framework. In G. Amati, C. Carpineto, and G. Romano, editors, *ECIR*, volume 4425 of *Lecture Notes in Computer Science*, pages 52–63. Springer, 2007.