New Utility Models for the Garnata Information Retrieval System at INEX'08

Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, Carlos Martín-Dancausa, and Alfonso E. Romero

Departamento de Ciencias de la Computación e Inteligencia Artificial
E.T.S.I. Informática y de Telecomunicación, Universidad de Granada, 18071 – Granada, Spain
{lci,jmfluna,jhg,cmdanca,aeromero}@decsai.ugr.es

Abstract. In this work we propose new utility models for the structured information retrieval system Garnata, and expose the results of our participation at INEX'08 in the AdHoc track using this system.

1 Introduction

Garnata [5] is a Structured Information Retrieval System for XML documents, based on probabilistic graphical models [8,9], developed by members of the research group "Uncertainty Treatment in Artificial Intelligence" at the University of Granada. Garnata has already been tested at two editions of the INEX Workshop [4,6], and its theoretical basis is explained in more detail in [1,2].

Garnata computes the relevance degree of each component or structural unit in a document by combining two different types of information. On the one hand, the specificity of the component with respect to the query: the more terms in the component appear in the query, the more relevant becomes the component, that is to say, the more clearly the component is only about (at least a part of) the topic of the query. On the other hand, the exhaustivity of the component with respect to the query: the more terms in the query match with terms in the component, the more relevant the component is, i.e., the more clearly the component comprises the topic of the query. The components that best satisfy the user information need expressed by means of the query should be, simultaneously, as specific and exhaustive as possible.

These two dimensions of the relevance of a component with respect to the query are calculated in a different way. To compute the specificity, the probability of relevance of each component is obtained through an inference process in a Bayesian network representing the structured document collection. The exhaustivity is obtained by first defining the utility of each component as a function of the proportion of the terms in the query that appear in this component. Then the Bayesian network is transformed into an influence diagram which computes the expected utility of each component, by combining the probabilities of relevance and the utilities in a principled way.

S. Geva, J. Kamps, and A. Trotman (Eds.): INEX 2008, LNCS 5631, pp. 39–45, 2009.

In this work we propose a modification of the system by defining the utility in a different manner, in such a way that those components that do not contain most of the query terms are penalized more heavily. By defining a parametric model, it is possible to adjust the degree of utility to make the system behave more similarly to a strict AND (if not all or almost all the query terms are in the considered component, this one will be scarcely relevant) or to a less strict AND.

2 Utility Models in the Garnata System

As we focus in this work on the utility component of the Garnata system, we will not enter into details of the Bayesian network model representing the document collection. This model is able to efficiently compute the posterior probabilities of relevance of all the structural units U of all the documents, given a query Q, $p(U|Q)$. These probabilities represent the specificity component of each structural unit U: the more terms indexing U also belong to Q, the more probable is U.

The Bayesian network is then enlarged by including decision variables R_U, representing the possible alternatives available to the decision maker (retrieve unit U), and utility variables V_U, thus transforming it into an influence diagram. The objective is to compute the expected utility of each decision given Q, $EU(R_U|Q)$.

In Garnata the utility value V_U of each structural unit U is made of a component which depends on the involved unit, other component which depends only on the kind of tag associated to that unit, and another component independent on the specific unit (these three components are multiplied in order to form the utility value, see [4]).

The part depending on the involved unit, which is the only one we are going to modify, is defined as the sum of the inverted document frequencies of those terms contained in U that also belong to the query Q, normalized by the sum of the idfs of the terms contained in the query: a unit U will be more useful (more exhaustive), with respect to a query Q, as more terms of Q also belong to U:

$$nidf_Q(U) = \frac{\sum_{T \in An(U) \cap Q} idf(T)}{\sum_{T \in Q} idf(T)} \tag{1}$$

$An(U)$ in the previous equation represents the set of terms contained (either directly or indirectly) in the structural unit U.

3 New Utility Models

As it can be observed from Eq. (1), the utility or exhaustivity of a structural unit U with respect to a query Q grows linearly with the number of query terms appearing in U (reaching a maximum equal to 1 when all the terms of the query appear in the unit). In our experience with the system in different applications [3,4], we have observed that this linear growing, when combined with the probabilities computed from the Bayesian network (which measure specificity), can cause that small structural units, which only match with a fraction of the query terms, become more relevant that other, greater structural units that contain more terms from the query. In many cases this behaviour is not the expected one, because probably a user who employs several terms to express his/her query is expecting to find most of these terms in the structural units obtained as the answer of the system to this query. For that reason we believe that it is interesting to define other utility models which give more importance (in a non-linear way) to the appearance of most of the terms in the query.

In this work we propose a parametric non-linear utility model that, as the parameter grows, the more terms from the query must be contained in a structural unit in order to get a high utility value for this unit. A way of obtaining this behaviour is through the use of the following transformation:

$$nidf_{Q,n}(U) = nidf_Q(U)\frac{e^{(nidf_Q(U))^n} - 1}{e - 1} \tag{2}$$

In this way, when $n = 0$ we have $nidf_{Q,0}(U) = nidf_Q(U)$, that is to say, we reproduce the original model, and the greater the value of the integer parameter n, we obtain a behaviour more similar to a strict AND operator. In Figure 1 we can observe several plots of the function $x\frac{e^{x^n}-1}{e-1}$ for different values of n.

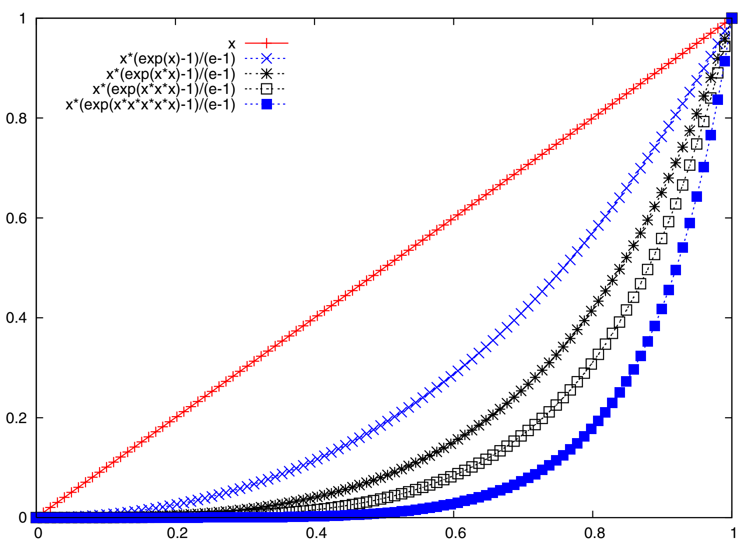

Fig. 1. Function $x\frac{e^{x^n}-1}{e-1}$, for $n = 0, 1, 2, 3, 5$

4 Experimental Results

In this INEX 2008 edition, we have participated submitting nine runs in the AdHoc track (content only). More specifically, three in each of the *Focused*, *Relevant in Context* and *Best in Context* sub-tasks. Table 1 shows the positions in the ranking according to the official evaluation measures ($MAgP$ for *Best in Context* and *Relevant in Context*, and $iP[0.01]$ for *Focused*), the sub-task and finally the run identifier.

Table 1. Runs submitted to the INEX'2008 AdHoc tasks and positions in the rankings

Position	Value	Sub-task	RunId
52	0.468856	Focused	p8_u3_exp_5_1110
53	0.467071	Focused	p8_u3_exp_3_1110
54	0.448733	Focused	p15_u3_exp_5_1110
25	0.158177	Relevant in Context	p8_u3_exp_5_1110
26	0.158177	Relevant in Context	p8_u3_exp_5_0100
27	0.152320	Relevant in Context	p8_u3_exp_3_1110
18	0.146799	Best in Context	p8_u3_exp_5_0100
19	0.146536	Best in Context	p8_u3_exp_3_0100
22	0.138141	Best in Context	p15_u3_exp_3_0100

Table 2. Importance of the different types of units used in the official runs

Tag	Weight file 8	Weight file 15
name	20	200
title	20	50
caption	10	30
collectionlink	10	30
emph2	10	30
emph3	10	30
conversionwarning	0	0
languagelink	0	0
template	0	0
default value	1	1

With respect to the parameters, we have used the weight files 8 and 15 (p8 and p15 as prefixes of the run identifiers), and utility file 3 (u3, contained in the identifiers), with the first values presented in Table 2 and in Table 3 the second ones (see [4] for details about these parameters and their use within the model). We have experimented with two values of the parameter n in Eq. (2), 3 and 5 (exp_3 and exp_5, also contained in the identifiers). These values were selected by means of experimentation with previous INEX collections. Finally, the suffix of the run identifier corresponds to the values of each of the four configurations of the component of the utility function independent on the involved unit (see [4]).

Table 3. Relative utility values of the different types of units used in the official runs

Tag	Utility file 3	Tag	Utility file 3
conversionwarning	0	section	1.25
name	0.85	p	1.5
title	0.85	body	2.0
collectionlink	0.75	emph2	1.0
languagelink	0.0	emph3	1.0
article	2.5	default value	1.0

Table 4. Comparison between runs with and without applying the transformation in Eq. (2)

With $nidf_Q(U)$	With $nidf_{Q,n}(U)$	%Change	Sub-tasks	Run Id.
0.366249	0.468856	28.01	Focused	p8_u3_exp_5_1110
0.366249	0.467071	27.53	Focused	p8_u3_exp_3_1110
0.341804	0.448733	31.28	Focused	p15_u3_exp_5_1110
0.083034	0.158177	90.50	Relevant in Context	p8_u3_exp_5_1110
0.067706	0.158177	133.62	Relevant in Context	p8_u3_exp_5_0100
0.083034	0.152320	83.44	Relevant in Context	p8_u3_exp_3_1110
0.075842	0.146799	93.56	Best in Context	p8_u3_exp_5_0100
0.075842	0.146536	93.21	Best in Context	p8_u3_exp_3_0100
0.078910	0.138141	75.06	Best in Context	p15_u3_exp_3_0100

Although there has been a significant reduction of runs submitted in this 2008 edition – measured as focused retrieval – (*Focused*: from 79 last year to 61 this edition; *Relevant in Context*: from 66 to 40; *Best in context*: from 71 to 35), we could say that in terms of the percentiles of the positions in the rankings, we are improving our results in *Relevant in Context* (from percentiles 66-74 last year, to 62-67 this year) and *Best in Context* (from 63-70 to 51-62), and keeps more or less the same positions in *Focused* (from 84-89 to 85-88).

It is noticeable that within the *Focused* task, Garnata's performance is relatively low, and keeps more or less the same positions than last year, and how the methods described in [4] for adjusting the output for the requirements of the other two tasks make a good job from the raw results generated by Garnata. Clearly *Best in Context* is the sub-task where the performance is higher, and where the best improvement is achieved.

In order to better determine the improvement obtained by the new utility model presented in this paper, we have run an experiment without using the transformation presented in Eq. (2), but applying instead the original Eq. (1), $nidf_Q(U)$. Table 4 shows the values of the official evaluation measures with the old utility model used in previous editions (first column), this year with the new model (second column) and the percentage of change (third column). As noticed, the percentages of change are generally quite large, and this fact confirms our initial hypothesis that the new transformation could improve the results.

We have carried out another series of experiments, motivated by the following fact: we realised that among the systems obtaining the best results in the official competition at INEX'08 [7], there are many systems that do not return any possible structural unit as a result but only some of them, typically only content-bearing elements like section, paragraphs or the complete article. In contrast, our official runs retrieved almost any elements, and this may be a source of poor behaviour specially when removing overlapping elements. So, we have repeated our official experiments but filtering the results in order to retrieve only article, or only article, body, section and paragraph elements. This can be easily done by using an utility file giving weight zero to all the structural units except the selected ones (with weight equal to one). The results of these experiments are displayed in Table 5.

Table 5. Runs retrieving only content-bearing elements and positions in the rankings

article+section+...		only article			
Position	Value	Position	Value	Sub-task	RunId
48	0.517808	52	0.482262	Focused	p8_u3_exp_5_1110
46	0.524948	52	0.478478	Focused	p8_u3_exp_3_1110
52	0.474641	54	0.455649	Focused	p15_u3_exp_5_1110
20	0.171119	27	0.157455	Relevant in Context	p8_u3_exp_5_1110
24	0.164420	27	0.157455	Relevant in Context	p8_u3_exp_5_0100
22	0.168308	27	0.155347	Relevant in Context	p8_u3_exp_3_1110
20	0.146501	14	0.168893	Best in Context	p8_u3_exp_5_0100
22	0.140705	14	0.167468	Best in Context	p8_u3_exp_3_0100
24	0.131170	18	0.148391	Best in Context	p15_u3_exp_3_0100

We can observe that this strategy of retrieving only the more general elements is useful for the *Focused* and *Relevant in Context* tasks, where we would obtain better positions in the ranking (going from percentiles 85-88 to 75-85 in *Focused* and from 62-67 to 50-60 in *Relevant in Context*, when using the four elements selected). However, the results are slightly worse for the *Best in Context* task (going from percentiles 51-63 to 57-68) in the case of using the four elements but better when using only the article element. These results point out that the choice of the structural elements to be retrieved has a non-negligible impact on the performance of an XML retrieval system.

5 Concluding Remarks

In this paper we have presented the participation of the University of Granada group in the 2008 INEX edition in the AdHoc tasks. This is based on the work developed in previous years, but introducing a new utility model which gives more importance (in a non-linear way) to the appearance of most of the terms in the query. We have shown in the previous section that this new approach considerably improves the results with respect to not using it.

With respect to the comparison of our results with the rest of participants, we could say that we are in the middle of the rankings, improving with respect to the last edition of INEX.

Regarding future research in the context of INEX, we have to work in the improvement of the raw results of Garnata, as they are the base for the different sub-tasks, and in the filtering strategy used to remove overlapping elements. Also, we have designed an approach to answer CAS queries, which will be evaluated in the next edition of the evaluation campaign.

Acknowledgments. This work has been jointly supported by the Spanish Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía, Ministerio de Ciencia de Innovación and the research programme Consolider Ingenio 2010, under projects TIC-276, TIN2008-06566-C04-01 and CSD2007-00018, respectively.

References

1. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: Using context information in structured document retrieval: An approach using Influence diagrams. Information Processing & Management 40(5), 829–847 (2004)
2. De Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: Improving the context-based influence diagram for structured retrieval. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 215–229. Springer, Heidelberg (2005)
3. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Martín, C., Romero, A.E.: An information retrieval system for parliamentary documents. In: Pourret, O., Naim, P., Marcot, B. (eds.) Bayesian Networks: A Practical Guide to Applications, pp. 203–223. Wiley, Chichester (2008)
4. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Martín-Dancausa, C.J., Romero, A.E.: The Garnata information retrieval system at INEX 2007. In: Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) INEX 2007. LNCS, vol. 4862, pp. 57–69. Springer, Heidelberg (2008)
5. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Romero, A.E.: Garnata: An information retrieval system for structured documents based on probabilistic graphical models. In: Proceedings of the Eleventh International Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), pp. 1024–1031 (2006)
6. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Romero, A.E.: Influence diagrams and structured retrieval: Garnata implementing the SID and CID models at INEX 2006. In: Fuhr, N., Lalmas, M., Trotman, A. (eds.) INEX 2006. LNCS, vol. 4518, pp. 165–177. Springer, Heidelberg (2007)
7. Geva, S., Kamps, J., Trotman, A. (eds.): INEX 2008 Workshop Pre-proceedings (2008)
8. Jensen, F.V.: Bayesian Networks and Decision Graphs. Springer, Heidelberg (2001)
9. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan and Kaufmann, San Mateo (1988)