Explaining the Relevance of Court Decisions to Laymen

Gwen R. WILDEBOER[a], Michel C.A. KLEIN[a], Elisabeth M. UIJTTENBROEK[b]
[a]*Department of Artificial Intelligence, Vrije Universiteit, The Netherlands*
[b]*Faculty of Law, Vrije Universiteit, The Netherlands*

Abstract. In the context of intelligent disclosure of case law, we report on our findings with respect to the presentation of relevant court decisions back to the laymen users. For this presentation we first localize the relevant legal concepts in the cases using shallow NLP techniques. Hereafter we investigated the use of techniques from the field of recommender systems, i.e. keyword style explanation and influence style explanation, to present the cases to the user in an understandable way. In order to find out if we succeeded in that respect, we conducted a small user satisfaction research. It shows promising results, and gives us some directions for future research.

Keywords. tort law, representation of legal texts, laymen, fingerprints, recommender systems, user satisfaction

1. Introduction

The judiciary is faced with enormous case loads. Alternative dispute resolution mechanisms such as mediation can help to reduce this workload. Mediation is not always preferred (if known at all), in particular since litigants are often not aware of their chances in court, and normally overestimate their chances. In the BEST-project[1] we aim at providing disputing parties with information about their legal position in a liability case. We are developing a system that supports users by retrieving relevant case law, i.e., court decisions. In this way parties are given the opportunity to form a judgment about whether they can hold another party liable for certain caused damage or if they can be held liable themselves. Also, parties are able to determine how much room for negotiation is available when settling the damage. By information about previous court decisions, where relevant taking into consideration other factors such as time, costs, emotions, etc., a well-rounded impression is obtained about a parties' BATNA (Best Alternative To a Negotiated Agreement), that is: the best option a party has if negotiation fails [1].

An important part of this project is the communication with the user. Communication happens at the start of the project, when the user has to provide the program with his case, and at the end of the program, when the user obtains an output

[1] www.best-project.nl

that should enable him[2] to establish his BATNA. The different phases in the BEST-project are graphically depicted below.

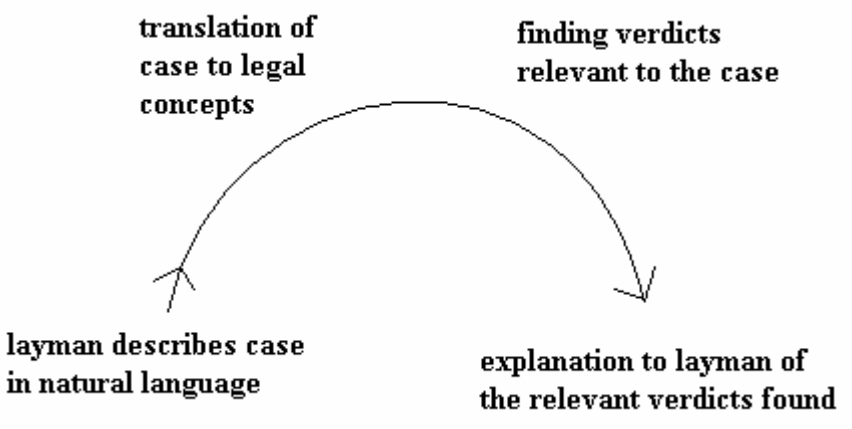

Figure 1. Overview of phases in the BEST-project.

In this paper, we focus on the last phase of the project, the explanation to the user of the relevant verdicts for a specific case. With this focus, we assume that the search part has already been conducted. For more information on the search part in the BEST project, see [2] and [3].

Many systems for legal information retrieval focus on legal practitioners. However, in our use case the user is assumed to be a layman which has serious difficulty with understanding legal concepts and reading legal texts. The explanation of the search results is therefore an important issue.

We report on an automatic method to present and explain relevant case law to a layman user in such a way that he:

1. understands the key concepts in the presented case better;
2. gains better insight in his own case.

This enhanced insight is in comparison to a user confronted with just the text of relevant verdicts. The research is concluded by a small user satisfaction study to find out whether the proposed presentation is indeed useful to prospective users.

To present an understandable explanation of the relevant verdicts, we take two steps. First, we localize the legal concepts that are relevant for the user's case in the verdicts and we decide which paragraphs are relevant for which concepts. Second, we present the user with an account of the verdict based on these relevant paragraphs and techniques from recommender systems.

In the research on which this paper is based, we try to answer the following research questions:

- Is it possible to locate the legal concepts in the verdicts, one way or the other?
- Which possibly useful presentations of these concepts can we identify?
- Which of those is most promising?

[2] 'He, him, his' can also always be read as 'she and her'.

- Is the chosen presentation good enough to indeed provide the user with insight into his position in his own case?
- If the presentation is not good enough yet, or as an addition, what other possibilities might be interesting for future research?

In the next section we will describe the techniques from recommender systems that inspired our approach. Section 3 will explain the localization of legal concepts. Section 4 will illustrate how the verdicts are eventually presented to the user. This presentation is evaluated with a user satisfaction study, on which we report in Section 5. The paper is finished with some conclusions and remarks for future research.

2. Using techniques from Recommender Systems

We have chosen to explore the possibilities of techniques from recommender systems to process the verdicts in a way that yields a useful presentation to the user.

Recommender Systems are usually divided into two approaches: Collaborative Filtering (CF) and Content Based Filtering (CBF). In Collaborative Filtering, the preferences of communities of similar users are used to decide on recommendations for the current user [4]. With Content Based Filtering the content of certain items is processed and based thereon a decision is made about whether the user will probably be interested in the item, based on some predefined user characteristics and a history of interest in earlier items [5].

For the task at hand, Content Based Filtering might very well be useful, as we want to process the verdicts on content to explain their relevance. The content of a paragraph decides whether the user will be interested in that paragraph. Of course, this is not based on the preferences of the user, but on the relevance of the legal content.

Another interesting prospect is that in Recommender Systems sometimes the recommendation is accompanied by a *reason*. An 'explanation mechanism' tries to explain why the program believes that the user will be interested in the prospective item [6]. It is interesting to investigate whether the techniques used to establish the reason for recommendation are also feasible for explaining to the user why those specific verdicts are represented to him as relevant for his case. A problem might be that in Recommender Systems the search for recommendable items is tied to the reasoning about why a certain recommendation was made. In this project, the search is done separately, and only afterwards it is tried to reestablish the reasons behind the selection of the final set of verdicts. History information about earlier recommendations is not available, so it can only be based on the contents of the verdicts under scrutiny at the moment.

Explanation systems have proven to be able to gain trust of users and thereby improve the acceptance of the recommender system, if the right form of explanation is chosen [7]. In our project, acceptance of the system is of course also vital, as otherwise the system will be useless. Added to that, issues of law are always significant to people and when they are personally involved with things at stake, they will not give their trust so easily. Judicial experts are usually held in esteem and their opinion will be of great value to laymen without any legal knowledge. However, as assisting software programs are still relatively new to people, they will not trust them quite so easily. Therefore, explanation systems might help in gaining trust.

Sometimes, the ways in which the effectiveness of explanation systems can be tested are divided in two approaches: the *promotion approach* and the *satisfaction*

approach. According to the promotion approach the best explanation system is the one that is most successful at convincing the user to follow the recommendations given. The satisfaction approach believes that the best explanation is the one that is best able to let the user assess the quality of the things that are recommended. We believe that for our own project, the truth is somewhere in between. On the one hand, the explanation should indeed convince the user that the verdicts he is given are relevant with respect to his own case. On the other hand it would be very useful if it enables him to assess the quality of the verdicts as to at which points they concur and at which points they conflict with his own case. Explanation systems that are interesting in this respect are Keyword Style Explanation and Influence Style Explanation [8]. In Keyword Style Explanation the user is given a table explaining which words in his profile and in the content of the item had the most influence on the rank of the item. This can possibly be applied in our project to the terms that are used to search the case law. In Influence Style Explanation, the system tells the user how their interactions with the recommender system influenced the recommendation. This could possibly be applied in our project by explaining how the key concepts in the original description of the user case influenced the selection of the verdicts.

For the application of both techniques, we first need to localize the legal concepts in the verdict, since they determine the 'relevant content' of the verdict. This localization is described in the next section.

3. Localizing legal concepts and relevant paragraphs

Concept-based search using thesauri is not a new technique in Information Retrieval. Most concept-based approaches nowadays are based on the use of document vectors (sometimes also called 'fingerprints') by Salton in 1989 [9], but also Bing described the advantages of this search method in 1987 [10]. The general idea of this approach is as follows: for every document a *document vector* is made; this is a list of the terms from the used thesaurus that are found in the document. Each term found is assigned a unique identifier and a relevance score that indicates the frequency of the term in the document and the specificity of it in the thesaurus. All this has to be done with an indexing algorithm. This algorithm first detects sentences in the document, then removes the stop words from these sentences, normalizes (stems) the remaining words and finally uses a specified thesaurus to identify words and phrases in the document [11]. A search can then be performed by calculating a 'document vector' for the search terms and comparing this vector with the vectors of the documents. In the BEST project, 36 search documents have been made that reflect a specific legal concept. By comparing the vector for these search documents with the vectors of the verdicts in the database, we identify which verdicts are relevant for a specific concept [3]. The thesaurus of the BEST-project for those 36 concepts (currently) consists of 1678 terms (words and phrases).

Since experiments have shown us that it is quite difficult to localize the legal concepts that are extracted from the user's case in a direct manner (e.g., by keyword search), we have decided to use the fingerprints from the search part of the project for localization. In GATE (General Architecture for Text Engineering) we first tokenize the relevant verdicts, then stem them, use a gazetteer to annotate words and phrases belonging to a concept, based on their fingerprint and finally use a transducer to be able to visualize the concepts belonging to the various annotations. We use the Snowball

stemmer, a flexible gazetteer in combination with the OFAI gazetteer and the JAPE transducer. The fingerprints of each concept are provided to GATE in lists of all corresponding terms and phrases, also in stemmed version. The result is an highlighting of the terms from the fingerprints in the verdicts. An example is shown in Figure 2.

In the next section we will describe how we process these annotations with

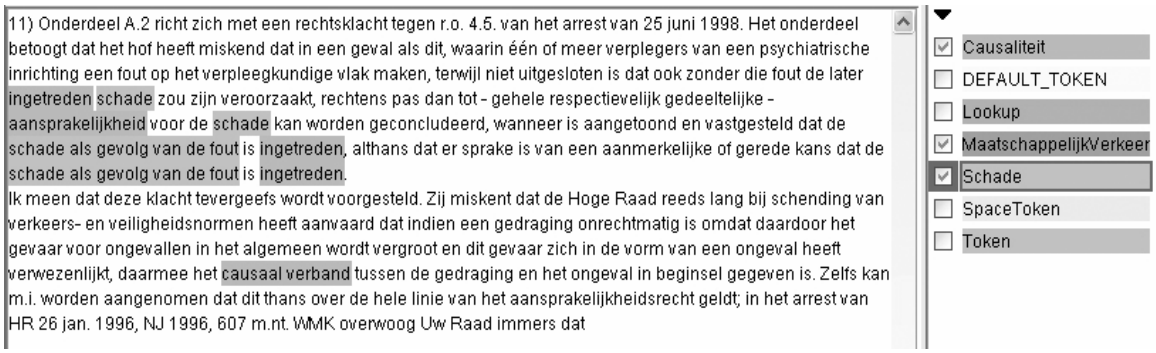

Figure 2. Example of annotation of terms for the concepts 'schade' (damage) and 'causaliteit' (causality).

Recommender techniques discussed earlier to arrive at the final presentation of the verdict to the user.

4. Output of the method

With the annotation of the terms from the fingerprints, we now determine which paragraphs are relevant for which legal concepts. However, it is not the case that a paragraph is always relevant if it contains a term from the search document. Some paragraphs will contain terms from multiple fingerprints, and other paragraphs will only contain a single, not very important term. To deal with this, we empirically design a number of rules that state how many times a term, or multiple terms from the same fingerprint, have to occur in a paragraph to consider that paragraph relevant for the particular concept. In these rules, we also take the weight of the term in the fingerprint into account. [3] When a paragraph is relevant, we highlight it entirely (so the highlighting of the separate terms disappears) and provide the paragraph with a comment that states the legal concept for which the paragraph is relevant.

Similar to *Keyword Style Explanation*, we list all legal concepts found in the verdict (corresponding to those extracted from the user case in another part of the program) and explain them in general wording at the top of the verdict.

Besides this Keyword Style Explanation, we also use *Influence Style Explanation*. Certain terms or concepts are used to link the verdict to the user case. If for example the verdict is about a 'traffic accident', then the user will be pointed to the similarities and differences in the relevant legal facts of their case This linking is done for multiple concepts in order to help the user apply certain aspects from the verdict to the user case.

[3] The details of the rules can be found in the report on which this paper is based: Wildeboer, G.R., 2007. Available from http://www.cs.vu.nl/~mcaklein/Wildeboer-2007.pdf.

The combination of these two methods results in our final output of this approach, which is depicted in Figure 3. For this study, we have manually added these explanations as comments to a document, but all explanations are based on a list of concepts that occur in the case description of the user and an implementable set of rules that use standard sentences. In a future version of the system, the pre-defined explanations will be automatically added to the document, based on the same set of rules.

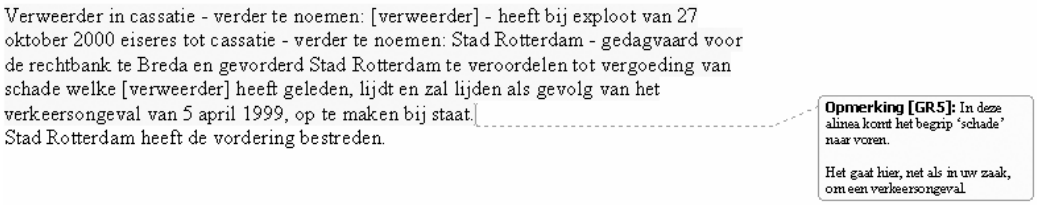

Figure 3. Example of relevant paragraph with Keyword Style and Influence Style Explanation

5. User study

We are very interested to see whether prospective users will really find our presentation of the relevant verdicts helpful in determining their BATNA. To investigate this we approached a group of 21 participants for a small study. As the group is relative small, and the selection is not a-select (the participants are taken from people in our social network), we realize that we cannot draw strong conclusions, but we think it will nevertheless shed some light on the value of our presentation.

None of the participants has any professional legal knowledge. The 21 people are divided into three groups of seven. Each group receives a general explanation of the study; a fictitious, but realistic, description of a case; four verdicts and three different types of questions.

The case is the following: *"A young cyclist with ill-functioning lights was hit by a car that was driving too fast. Because of the accident, one foot was broken and another had to be amputated. The victim is not able to play sports anymore, and he also failed for his high school exam, which he attributes to the depression following the accident. Already before the accident he has had depression related complaints, but those were not present any more at the moment of the accident. He wants to have a compensation for the medical costs and his depression."*

The difference between the groups is the extra information given with the verdicts. Group 1 just receives the verdicts, with no explanation whatsoever. For Group 2 the verdicts are processed according to the Keyword Style Explanation explained in the previous section. Group 3 gets the verdicts processed with Keyword Style and Influence Style Explanation.

Apart from some personal information that is collected to be able to account for differences in age, education and legal knowledge, there are three different types of questions. The first category consists of 'subjective' questions; propositions with an answering scale from 1 (I do not agree at all) to 5 (I agree completely). These are designed to measure the confidence the user has in the program and the extent to which

they feel the program is useful to establish their BATNA. The questions in this category are:

1. By now I know what would probably be the outcome of this case if it would be bought to court.
2. I have a better insight in what are the most important aspects of my case.
3. Because of my insight in the case I am able to judge whether an offer to settle the case outside the court would be beneficial.
4. If I would have a real case I would invest time in this system to assess my legal position.
5. I trust the relevance of the verdicts returned by the system.
6. I trust the correctness of the explanation about the relevant aspects in my case.

The questions in the second category, i.e. the 'objective' questions, are in a kind of exam style. Those questions are designed to test the knowledge of the user about the provided case, the legal concepts and their BATNA based on what they learn from the verdicts. The last category contains some open questions in which the users can express what they liked about the program, what they miss and anything else they want to share. These questions are:

1. Describe three important aspects in your case.
2. Would the judge probably conclude that the damage to the feet and the costs that had to be made are a consequence of the accident?
3. Would the judge probably conclude that the depression and the costs that had to be made are a consequence of the accident?
4. Is it likely that the judge will grant compensation for the fact that the victim isn't able to play sports anymore?
5. Is it likely that the judge will grant the full compensation that is requested?
6. [Offer for settlement] Would you accept this offer?

Our overall hypothesis is that group 3 will have the highest average scores for the subjective questions (meaning the highest confidence in and satisfaction with the program) and perform best on the objective questions (which means that they will answer most questions correctly). Group 2 will have lower scores than group 3 and group 1 will have the lowest scores. This because we believe that the extra information provided to group 2 and group 3 will help them with understanding the verdicts and the implications for their own case. This extra information is expected to enhance confidence and also to make it easier to establish the BATNA. Apart from these measures on the questions, we hypothesize that group 1 will probably need more time to complete the whole survey, as they will have to read the verdict on their own to find out what is relevant, whereas the other groups have the relevant paragraphs highlighted already. Out of the 21 surveys sent, we got 15 back; 5 in each group coincidently.

Table 1. Time needed to complete the survey

Respondent →	**1**	**2**	**3**	**4**	**5**	**Average**	**St. dev**
Group 1	100 min	55 min	60 min	100 min	50 min	73 min	24,9
Group 2	60 min	40 min	55 min	45 min	60 min	52 min	9,1
Group 3	30 min	35 min	20 min	40 min	35 min	32 min	7,6

Table 2. Average scores on the subjective questions (scale 1-5)

Question →	**1**	**2**	**3**	**4**	**5**	**6**	**Average**
Group 1	2,4	2,8	2,6	3	3,8	[4]	2,9
Group 2	3	3,2	3,0	3,8	3,2	3,4	3,3
Group 3	3,4	3,8	3,4	4	3,8	3,6	3,7

The results are presented in Table 1 and 2. Table 1 shows the time that the participants needed to complete the survey, for all 15 participants. The ranking in the average scores is as we hypothesized, and the relatively small standard deviation supports our confidence in the ranking. Table 2 shows the average scores on the five subjective questions per group of five respondents. This table also shows the expected ranking: group 3 is most confident and satisfied, while group 1 is least satisfied.

The objective questions have answers in free text, so we did no statistical analysis of the answers. Still it is possible to note that there are some interesting differences between the groups. As is predicted beforehand, none of the respondents in group 1 mentions the three legal concepts that were relevant in the user case 'damage, 'causality and 'act or omission violating an unwritten rule pertaining proper social conduct', where most of those in group 2 and 3 did. Further, all respondents believe that a judge will grant the victim full compensation of the medical expenses for his foot. The majority of those in group 1 and 2 believe that the judge will not grant expenses made because of the depression. The reason given for this belief is that the victim has had depressions before, so the causal relationship can not be established in their eyes. In group 3 there are remarkably more respondents believing that the depression-related expenses *will* be granted, which is the correct answer and therefore concurs with our hypothesis that group 3 performs best on the objective questions. Answers to the question whether a compensation for not being able to play sports anymore will be granted are rather varying. Main reason for this is the need of more detailed information about sports history of the victim and alternative career prospects. Almost none of the respondents believe that a judge will grant all claimed damages: from the reactions it seems they just assume that a judge will never give you exactly what you ask for. Finally, group 3 is more reluctant to take the offer of a settlement than the other groups (4 of the 6 wouldn't take the offer, whereas in groups 1 and 2 only 2 of the 6 would not take it).

The responses to the open questions might even be the most useful for our research, but also for the project as a whole, because the participants have been very honest. We even got phone calls from those wondering whether the task was really that difficult or it was just them. Multiple reactions are of the tenor 'please let me hire a lawyer' and 'this is way too difficult for a layman like me'. However, these remarks are in fact contradictory to the answers at the subjective questions. Apart from group 1, the average was above 'neutral' towards the positive side of the scale. This indicates that

[4] This question was about the extra information provided. Group 1 didn't get any extra information; hence this question wasn't relevant to that group.

they *did* learn something from the program (as could also be seen with the open questions), although they think it was too difficult for them.

Taking all the results together, we think we can be cautiously optimistic. As our final goal is to present the users with something like the version group 3 got, we will base our conclusion thereupon. This group is positive about their gained understanding of their case, and most of them have answered the objective questions in the way we envisioned beforehand. However, there are also some critical remarks, even from group 3, so we are not there yet. The verdicts are still very hard to read because of the legal jargon, but that is not something we see a solution for right now, as the verdicts will be the output of the system.

In short, our user study shows that the system helps users to gain understanding, but that users themselves still consider the interpretation as quite difficult. At the same time this is a justification for our whole research: the output of the system is very important to the overall success and therefore is worthy of thorough research.

6. Summary and Future Work

In this paper we have presented a method for presenting relevant court decisions to layman in such a way that they acquire a better understanding than when they just see the plain text of the verdict. The method has been partly implemented; another part has been executed manually following precisely described rules.

An important step in the method is the selection of relevant paragraphs in the text of the verdicts. In our case the relevance is determined by the presence of a (number of) legal concept(s) in a paragraph. Our implementation shows that this can be done automatically using shallow NLP techniques (i.e. stemming) and the weights of the terms in the document vectors that are used in the search process.

In addition, the user study has shown that the highlighting of relevant paragraphs reduces the time that a layman needs to "work through" the case, without sacrificing the understanding that he gains (if accompanied with some additional explanation[5]). This suggests that in some cases the highlighting of important paragraphs can be used as an alternative to summarization, as the parsing time reduces while no information is left out.

Moreover, the user study shows that an automatically generated explanation can help to improve the understanding of a case in less time. This holds both for the subjective understanding (whether people *think* that they understand it) and the objective understanding (whether they indeed draw the right conclusions). A general, keyword style based explanation already helps, and an influence style based explanation (an explanation related to the case that the user has to interpret) helps even more.

Finally, we have seen that this knowledge is used in the interpretation of the case that was presented to participants. It is apparent in their answers to the questions in the small survey that they have a more correct estimation of the chances in court.

Summarizing, taking into account the previously mentioned disclaimer about the small size of the study, we think that we may conclude that our method indeed helps to

[5] We did not test whether the understanding would still have been as good if no additional explanation was provided, but because the second group only got generic information, we think that the would be the case.

understand case law and also helps to improve the assessment of user's own legal situation.

In the future we will work on an implementation of parts that were not yet automated. We will also try to extend and improve the explanations. An important lesson is that effectively linking an explanation to a user's case requires identifying aspects that distinguish between different possible outcomes of a case. For example, in a traffic accident it is important whether one of the parties is a cyclist. Because the explanation is based on concepts that occur in paragraphs, we need concepts that reflect these aspects. Those concepts are often different than the concepts defined in the code. Future work in our project on searching case law shall have to take this perspective into account.

Acknowledgements

The BEST-project is funded under grant number 634.000.436 by the Netherlands Organisation for Scientific Research. The authors would like to thank all volunteers that participated in the user study and the Dutch Council for the Judiciary for using their database of case law.

References

[1] Berend R. de Vries, Ronald Leenes, and John Zeleznikow. Fundamentals of providing negotiation support online: the need for developing batnas. In John Zeleznikow and Arno R. Lodder, editors, *Second international ODR Workshop (odrworkshop.info)*. Wolf Legal Publishers, 2005.

[2] Michel C.A. Klein, Wouter van Steenbergen, Elisabeth M. Uijttenbroek, Arno R. Lodder, Frank van Harmelen. Thesaurus-based Retrieval of Case Law Proceedings of JURIX 2006, Paris, 7-9th December 2006.

[3] Elisabeth M. Uijttenbroek, Michel C.A. Klein, Arno, R. Lodder, Frank van Harmelen. Case law retrieval by concept search and visualization. Proceedings of ICAIL 2007, Stanford, California, 4-8th June 2007.

[4] O'Donovan, J. and B. Smyth (2005). Trust in recommender systems. International Conference on Intelligent User Interfaces, San Diego, ACM Press.

[5] Van Meteren, R. and M. Van Someren (2000). Using Content-Based Filtering for Recommendation. MLnet/ECML2000 Workshop, Barcelona.

[6] McSherry, D. (2005). "Explanation in Recommender Systems." Artificial Intelligence Review **24**(2): 179-197.

[7] Johnson, H. and P. Johnson (1993). "Explanation Facilities and Interactive Systems." Intelligent User Interfaces.: 159-166.

[8] Bilgic, M. (2004). Explanation for Recommender Systems: Satisfaction vs. Promotion. Computer Sciences. Austin, University of Texas. **Undergraduate Honors:** 27.

[9] Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of information by computer. Reading, Addison-Wesley.

[10] Bing, J. (1987). Designing text retrieval systems for "conceptual searching". International Conference on Artificial Intelligence and Law, Boston.

[11] van den Berg, J., C.C. van der Eijk, et al. (2004). "Constructing an Associative Concept Space for Literature-Based Discovery." Journal of the American Society for Information Science and Technology. **55**(5): 436-444.

[12] Herlocker, J.L., J.A. Konstan, et al. (2000). Explaining Collaborative Filtering Recommendations. ACM 2000 Conference on Computer Supported Cooperative Work, www.grouplens.org.