CASE LAW RETRIEVAL: CRITICAL EVALUATION OF CZECH LEGAL INFORMATION RETRIEVAL SYSTEMS

Jakub Harašta

Assistant Professor, Masaryk University, Faculty of Law, Institute of Law and Technology
Veveří 70, 625 00 Brno, CZ
jakub.harasta@law.muni.cz; http://cyber.law.muni.cz

Abstract: *This paper is concerned with a critical evaluation of three widely used Czech legal information retrieval systems: ASPI, Beck-online and Codexis. The legal information retrieved from these systems shape the existing Czech legal practice. Thus, the issue of how reliable their results are when compared to each other is of utmost importance. The paper reports results of an experiment aiming to frame the discussion on information retrieval from the Czech systems. The experiment uses relation-based searches to create three sets of court decisions. The union of these sets establishes ideal databases for individual research tasks that allow for a comparison of the Recall and the Precision of the individual systems. To assess relevance of documents in a system-independent way, three lawyers (two experts and a non-expert) provided for a binary coding of the relevance of the individual court decisions.*

1. Introduction

Lawyers were early users of electronic information retrieval (IR) systems. Retrieving information from electronic IR systems immediately became subject of discussion in literature. Issues of developing retrieval strategies and maintaining user competence [Bing 1978], or evaluation of relevance of retrieved documents for users' queries [Blair/Maron 1985] framed their expansion across the world of law. Understandably, retrieval of relevant documents (acts, case law, and literature) is not subject to a purely scholarly interest. A diligent search is important for ensuring clients' compliance with the law or for properly defending their interests in court. A practitioner is liable for incorrect legal advice, which might be caused by sloppy legal research. Therefore, an awareness of limits of specific information systems and their content is expected. Unfortunately, there is no literature focused on the evaluation of electronic IR systems in the Czech Republic. So far, authors and lecturers (including myself) have focused on tutoring students and judges to enhance literacy in the use of electronic IR.

There are three dominant IR systems in use in the Czech Republic – ASPI, operated by the Czech branch of Wolters Kluwer, Beck-online, operated by the Czech branch of C.H. Beck Verlag, and Codexis operated by the Czech-based Atlas Consulting. Over the past, it has become part of the canon to treat these systems as equal, as they contain similar (to a large extent) collections of case law. So far, no attention was paid to these systems in terms of critical evaluation of their results and, more importantly, to a comparison of these results [Gerson 1999; Knapp/Willey 2013; Mart 2013].

This paper aims to fix this issue by directly comparing results of systematic querying of ASPI, Beck-online, and Codexis. I retrieved case law related to individual Sections of Act no. 121/2000 on Copyright (Copyright Act, CA) and to Act no. 101/2000 on Personal Data Protection (APDP). A set of retrieved pairs «Section-Document» (*S-D*) for each Act is used to prepare an ideal subset to allow for a comparison between the IR systems. Additionally, to provide for a system-independent relevance assessment, I invited three lawyers to label the retrieved pairs *S-D* as either relevant or non-relevant for their understanding of the given Section.

After focusing on the related work in Section 2 of this paper, I thoroughly describe the experiment in Section 3. Section 4 describes results of the experiment in the form of recall and precision rates for the individual systems. Section 5 discusses the results offering their interpretation and pointing out possible problems with the used method. Section 6 concludes this paper.

2. Related Work

2.1. Information Retrieval

In general, performance of IR systems is measured in terms of Precision, Recall, F1 Measure, and Elusion [Ashley 2017, 222] of relevant documents. The precision is the ratio of the number of relevant documents retrieved out of the number of all documents retrieved. The recall is the ratio of the number of relevant documents retrieved out of the number of documents in the corpus. The F1 measure is a mean of precision and recall, where both measures are treated equally. The elusion is the proportion relevant documents that remain unretrieved. [Bing 1978, 384–385; Ashley 2017, 222].

The so called STAIRS study published by Blair and Maron in 1985 [Blair/Maron 1985] introduced lawyers to these concepts by pointing out an often insufficient recall rate for full-text retrieval by use of Boolean Searches. In this study, legal researchers achieved a recall rate of less than 20% even when they believed to retrieve more than 75% of relevant documents. Although the study focused on a discovery of documents containing and establishing facts of the case, and not on a retrieval of the case law relevant for legal argumentation, it has demonstrated limits of systems based solely on full-text queries. Subsequent studies by Dabney showed that similar limitations apply to other systems and, as such, are typical for full-text queries through Boolean Searches [Dabney 1986b]. Notably, some of the limitations shown by the STAIRS study were later empirically explained as typical for high recall searches [Sormunen 2001].

Boolean Searches are not the only way to retrieve relevant documents. The legal IR is currently divided between Boolean Searches, knowledge engineering (KE) frameworks, and NLP-based IR [Maxwell/schaffer 2008]. Understandably, all of the IR methods have advantages and disadvantages, as a perfect F1 measure (a perfect recall and a perfect precision) remains unachievable. The KE approach attempts to capture the way legal experts remember the cases, and express it in computer algorithms and data structures [Hafner 1987]. However, a significant drawback is in its problematic practical use. As pointed out by Maxwell and Schafer, it is domain-focused and often conducted on small, highly structured data collections. Moreover, the KE solutions are difficult to implement and scale [Maxwell/schaffer 2008, 66]. The NLP-based IR presents significant benefits in its scalability. It has also shown better results from its early days [Turtle 1994]. Nowadays, it is slowly appearing throughout the IR systems in some way or another. The ultimate goal seems to be to allow conducting legal IR as question answering [Maxwell/schaffer 2008, 68] or to allow a conceptual IR [Grabmair/Ashley/Chen/Preethi/Nyberg/Walker 2015].

2.2. Relevance in Information Retrieval

As mentioned in Subsection 2.1, the performance of IR is measured by retrieving a document relevant to a user query. However, the term «relevance» goes all too often undefined as we rely on its intuitive understanding. Only in 2017, Opijnen and Santos published a conceptualization of the term, where they derived the basic conceptual framework from general (non-legal) IR [Opijnen/Santos 2017]. This work distinguishes between different domains of relevance: Algorithmic Relevance, Bibliographical Relevance, Topical Relevance, Cognitive Relevance, Situational Relevance and Domain Relevance [Opijnen/Santos 2017, 73]. Three of these relevance types are used in the experiment described in Section 3 and will be described in further detail.

The Topical Relevance is the relation between the topic as formulated in the user request and the topic of the information objects. Relating a document to a specific topic is explicit or implicit by mapping and indexing terms, manual indexing of documents, a semi-automated classification and a relation-based search

[OPIJNEN/SANTOS 2017, 74]. The Cognitive Relevance is the relevance of information objects for individual user. As such, it is subjective and dependant on user's prior knowledge of the issue [OPIJNEN/SANTOS 2017, 81].

The Domain Relevance is the relevance of information objects within the legal domain itself. As such, it is independent of an individual query and carries with it information about the overall importance of the specific types of documents in the legal domain. Within the IR systems, it is the legal importance of classes of information objects or legal importance of individual information objects. [OPIJNEN/SANTOS 2017, 82–83].

3. Experiment Design

3.1. General Framework and Research Questions

In 1978, BING wrote that research strategies containing Boolean Searches are easy to learn but they are not easy to use [BING 1978, 380]. Another issue he mentioned is maintaining the research competence [BING 1978, 381]. From my own tutoring experience at the Faculty of Law of Masaryk University and at the Czech Judicial Academy, this has barely changed. The IR systems available today are more user friendly, but Boolean Searches remain difficult to use properly [VAN NOORTWIJK 2017, 10–11]. Achieving high recall and precision is difficult without proper training, and even then, the skill tends to deteriorate if not exercised on a regular basis. Additionally, the Czech off-the-shelf IR systems currently do not use KE frameworks or NLP-based retrieval.

A relation-based search is, similarly to other countries [OPIJNEN/SANTOS 2017, 74–75], very popular and is the very basic search technique for retrieving topically relevant documents. The relation-based search, as part of the Topical Relevance [OPIJNEN/SANTOS 2017, 74], allows retrieval of documents based on their metadata, tying the document to a specific Section of a specific Act. This shifts the cognitive load, otherwise very high when using Boolean Searches to query non-intermediated legal information, away from the user. The user then queries the intermediated legal information pre-processed by the IR system operator. There is nothing objective about these relations. The providers of the individual systems create relations between individual documents or their parts (in this research, between a section and a court decision). Given this, differences are understandable. However, there are currently no data to support this claim or to evaluate the extent of differences.

Hence, the first research question: *Does the use of a specific IR system lead to different case law retrieved by the relation-based search?*

This is the typical way of retrieving legal information in the Czech Republic, and probably elsewhere as well. The lawyer determines applicable provisions of the law and then uses the relation-based query to retrieve related documents – related legal acts, commentaries, handbooks, or, as in the following experiment, the case law. Additionally, lawyers pay attention to whether or not the case was «published». Generally, a case law published in official court collections is considered more important than its non-published counterpart. There are three official court collections in the Czech Republic, one for each apex court – the Supreme Court publishes a collection titled *Sbírka rozhodnutí a stanovisek Nejvyššího soudu* based on Art. 24 of Act no. 6/2002 Sb., on Courts and Judges; the Supreme Administrative Court publishes a collection titled *Sbírka rozhodnutí Nejvyššího správního soudu* based on Art. 22 of Act. no. 150/2002 Sb., Code of Administrative Justice. Finally, the Constitutional Court publishes a collection titled *Sbírka nálezů a usnesení Ústavního soudu* based on Art. 59 of Act no. 182/1993 Sb., on Constitutional Court. As the information about publication in a court collection originates outside of individual IR systems, the class of published (therefore, more relevant under Domain Relevance [OPIJNEN/SANTOS 2017, 82–83]) cases is the same for each IR system.

Hence, the second research question: *Does the use of a specific IR system lead to different published case law retrieved by the relation-based search?*

3.2. Retrieval of Case Law

For the purpose of this experiment, I analysed the Copyright Act (CA) and the Act on Personal Data Protection (APDP). These Acts were not selected at random. My research and teaching often revolves around these Acts and I wanted to ensure further use of information collected for this experiment. I retrieved the case law by manually querying ASPI, Beck-online and Codexis for a case law related to individual Sections of the selected Acts. I queried ASPI on 10th October 2018 for the CA, and on 12th December 2018 for the APDP, Beck-online on 10th October 2018 (CA) and on 18th December 2018 (APDP), and Codexis on 11th October 2018 (CA) and 18th December 2018 (APDP). Only decisions of Czech courts are included in this experiment.

The retrieved case law is organised into Section-Document (*S-D*) pairs. In each pair, *D* is the docket number of the court decision retrieved by querying related documents to the specific Section of the CA or the APDP. *S* is the identification of the specific Section of CA or APDP. Organising the retrieved case law in this way allows accounting for a situation where the court decision relates to multiple sections of the Act, because e.g. multiple issues arise during the court proceeding.

Each *S-D* pair was coded (1 or 0) in a relation to the appearance of pair in the individual IR system. Also, each *S-D* pair was coded (1 or 0) as to whether the document *D* contained in the pair was published in one of the official court collections mentioned in Subsection 3.1.

Subsequently, I organised the retrieved pairs in a database as shown in Table 1. The first row reads as *S* being Article 2 of the CA and *D* being the decision docket no. 10 A 186/2013-49. This pair could be retrieved from ASPI, but not from Beck-online or Codexis (because the document *D* is not indexed as related to the section *S*), and the decision itself is not published in any of the official court collections. The second row reads as *S* being Article 21 of the APDP and *D* being the decision docket no. I. ÚS 1783/10. This pair appears in ASPI and Codexis, but not in Beck-online, and the decision itself is published in one of the official court collections.

I use this database to calculate the recall rates of the individual IR systems, as is elaborated in Section 4.

Section	Act	Docket no.	ASPI	Beck-online	Codexis	PUB
2	CA	10 A 186/2013-49	1	0	0	0
21	APDP	I. ÚS 1783/10	1	0	1	1

Table 1: Example of two rows from the database of retrieved and further coded *S-D* pairs

Section	Act	Docket no.	Expert (CA)	Expert (APDP)	Non-expert
2	CA	10 A 186/2013-49	0		1
21	APDP	I. ÚS 1783/10		1	1

Table 2: Example of two rows from the database of *S-D* pairs with manually coded relevance of a court decision for human understanding of the Section

3.3. Manual Selection of Relevant Pairs

The second part of the experiment involves a human assessment of relevance of the retrieved *S-D* pairs. Three lawyers were engaged in this part of the experiment. The first lawyer served, due to several years of research and practical experience, as a domain expert for data protection and evaluated relevance of the pairs *S-D* retrieved for the APDP. The second lawyer served, due to his research and professional engagement in technology transfer, as a domain expert for intellectual property and evaluated relevance of the pairs *S-D* retrieved for the CA. The third lawyer served as a non-expert for both the APDP and the CA.

OPIJNEN and SANTOS noted that relevance has a contextual dependency, because it is always measured to the matter at hand – it is always related to a specific issue [OPIJNEN/SANTOS 2017, 70]. Therefore, three participating lawyers were presented with the retrieved pairs *S-D* and were asked to code (1 or 0) whether the presented document *D* helped in their understanding of the section *S*.

Subsequently, I organised the retrieved pairs in a database shown in Table 2. The first row reads as *S* being Article 2 of the CA and *D* being the decision docket no. 10 A 186/2013-49. This pair was coded as non-relevant for the expert's understanding of Section 2 of the CA and relevant for the non-expert's understanding of Section 2 of the CA. The second row reads as *S* being Article 21 of the APDP and *D* being the decision docket no. I. ÚS 1783/10. This pair was coded as relevant for both the expert's and the non-expert's understanding of Section 21 of the APDP.

4. Results

4.1. General Results

The total number of the *S-D* pairs retrieved by the relation-based search across all three IR systems is 403 for the CA and 262 for the APDP. The total number of the *S-D* pairs retrieved by the relation-based search across all three IR systems, where the document *D* was published in one of the courts official collections, is 92 for the CA and 68 for the APDP. Overall, the set of retrieved pairs contains 193 court decisions (26 published in the court collections) related to the CA, and 154 court decisions (37 published in the court collections) related to the APDP.

The total number of the *S-D* pairs deemed relevant by the experts was 107 for the CA and 130 for the APDP. The total number of the *S-D* pairs deemed relevant by the non-expert was 367 for the CA and 233 for the APDP. The total amount of the *S-D* pairs containing a published court decision deemed relevant by the experts was 35 for the CA and 40 for the APDP. The total amount of the *S-D* pairs containing a published court decision deemed relevant by the non-expert was 88 for the CA and 57 for the APDP.

To allow for a comparison of the three IR systems on the same terms, it is necessary to create an ideal database on top of what is essentially a union of their results. I prepared four ideal databases:

- a database containing all *S-D* pairs, where *S* is a section of the CA;
- a database containing all *S-D* pairs, where *S* is a section of the CA and *D* is a published court decision;
- a database containing all *S-D* pairs, where *S* is a section of the APDP;
- and finally a database containing all *S-D* pairs, where *S* is a section of the APDP and *D* is a published court decision.

Properties of these ideal databases, namely the total amount of the *S-D* pairs contained and the amount of the relevant pairs and the non-relevant pairs per both the expert and the non-expert assessment, are listed in Table 3.

Ideal database	Total	Expert		Non-expert	
		Relevant	Non-relevant	Relevant	Non-relevant
CA (all)	403	107	296	367	36
CA (published)	92	35	57	88	4
APDP (all)	262	130	132	233	29
APDP (published)	68	40	28	57	11

Table 3: Properties of the ideal databases used for the comparison of the three Czech IR systems. The number denotes the amount of the *S-D* pairs.

4.2. System-specific Results

In this part, I report the system-specific results for each of the three dominant IR systems. The results are reported as the Precision, the Recall and the F1 Measure achieved by ASPI, Beck-online and Codexis when querying four distinct ideal databases (comparative bases of documents) as described in Subsection 4.1. This allows for comparable results.

The Precision was computed as (TP)/(TP+FP), the Recall as (TP)/(TP+FN) and the F1 Measure as (2*TP)/(2*TP+FN+FP), where TP stands for the number of true positives, TN stands for the number of true negatives, FP stands for the number of false positives and FN stands for the number of false negatives.

The results are divided into two tables. Table 4 shows the Precision, the Recall and the F1 Measure for retrieving documents marked as relevant by the domain experts. Table 5 shows the Precision, the Recall and the F1 measure for retrieving documents marked as relevant by the non-expert lawyer.

Ideal database	ASPI			Beck-online			Codexis		
	P	R	F1	P	R	F1	P	R	F1
CA (all)	.254	.729	.377	**.459**	.579	**.512**	.274	.635	.383
CA (published)	.426	.829	.563	**.467**	.800	**.589**	.451	.657	.535
APDP (all)	.534	.723	**.614**	**.580**	.362	.445	.507	.546	.526
APDP (published)	.623	.825	.710	.638	.750	.690	**.640**	.800	**.711**

Table 4: The Precision, the Recall and the F1 Measure for retrieving documents marked as relevant by the domain-experts. The maximal values are in bold

Ideal database	ASPI			Beck-online			Codexis		
	P	R	F1	P	R	F1	P	R	F1
CA (all)	.915	**.766**	**.834**	.941	.346	.506	**.948**	.640	.764
CA (published)	.971	.750	**.846**	.967	.659	.784	**.980**	.568	.719
APDP (all)	.926	**.700**	**.797**	**.938**	.326	.484	.843	.506	.633
APDP (published)	.887	.825	**.855**	.936	.772	.846	.860	.754	.804

Table 5: The Precision, the Recall and the F1 Measure for retrieving documents marked as relevant by non-expert lawyer. The maximal values are in bold

5. Discussion

When lecturing on legal IR, I often encounter people asking what system I would recommend for conducting legal research. Unfortunately, this paper cannot answer this question comprehensively. However, given the experiment described above and its resulting data, I can discuss several implications.

First of all, per my research questions formulated in Subsection 3.1.: Apparently, when relying to a relation-based search, different systems yield different results. This was expected. The relation-based search queries relations created by individual providers of IR systems. Results are, inherently, system-specific, and cannot be objective. Therefore, using a specific IR system leads to both retrieval of specific set of case law, and a retrieval of specific set of published case law. Systems may contain mostly the same documents, but relations between those documents lead to different outcomes of any conducted legal research.

For the further discussion, I would like to draw attention to the relevance assessment conducted by two domain-experts and one non-expert lawyer. The Cognitive Relevance described by OPIJNEN and SANTOS

[OPIJNEN/SANTOS 2017, 80–81] is user-dependent. As such, its results also depend on prior knowledge of the lawyers engaged in the legal research. My distinction between a domain-expert and a non-expert lawyer was motivated by an assumption that a non-expert lawyer might assess the relevance differently, given the limited prior knowledge of the issue. Indeed, the non-expert lawyer assessed more documents as relevant, which led to a significantly higher Precision (and F1 Measure as a result) reported in Table 5. The non-expert lawyer consumed case law less discriminately than the domain-experts did.

This fact allows interpreting the results on two different levels – with two different types of users in mind. The first type of user is someone rather new to the selected specific issue, such as data protection or intellectual property. Such a user may greatly benefit from using ASPI that scored reasonably well in both the Recall and the Precision (the F1 Measure between .797 and .855). Both Beck-online (the F1 Measure between .484 and .846) and Codexis (the F1 measure between .633 and .804) scored lower. This could mean that ASPI indexes relations between sections and documents less strictly – and when a user approaches the system with little prior knowledge, this is beneficial.

On the other hand, the second type of user is someone who is a domain-expert and needs to further her knowledge or gain new insight into an issue. ASPI still offers the best Recall (.723 to .829) compared to other systems (.362 to .800 for Beck-online, and .546 to .800 for Codexis), but at significant (and understandable) cost in the Precision. The differences are noticeable and Beck-online seems a reasonable choice for an expert user. Its Precision is higher compared to the other two systems (.459 to .638), although not significantly (.254 to .623 for ASPI, and .274 to .640 for Codexis). The provider of Beck-online seems to be more careful when selecting relations to appear in the system – and the selection mimics both of the expert users quite well in their own relevance assessments.

Unfortunately, this study is far from conclusive. Given a small sample of documents (only two Acts), it cannot offer anything beyond estimated tendencies. A follow-up study in this area on a larger sample is needed. Additionally, the relevance assessment by the domain-experts and the non-expert lawyer is relatively simple as well. A scale would reflect user needs more accurately, compared to binary classification.

6. Conclusion

In this paper, I investigated whether a set of retrieved case law depends on the use of a specific IR system. I retrieved the case law from three individual IR systems, organised it into a database and created ideal databases to allow for a comparison of the results. Therefore, I can reliably conclude that outcomes of legal research relying on relation-based search are system-specific.

To my knowledge, this study is a first attempt to compare the three dominant Czech legal IR systems – ASPI, Beck-online, and Codexis. Therefore, even when suffering from shortcomings such as a small sample size and a significant simplification of human relevance assessment, I believe it is putting scientific rigor where it belongs, but where it is not usually readily available.

Legal research is a significant part of both academia and legal practice. As the number of documents grows, it is not, and never again will be, viable to rely on printed documents. We rely on IR systems and, in turn, these IR systems shape our legal knowledge. It is only logical to subject these systems to rigorous comparative and analytical efforts.

7. Acknowledgment and conflict of interest

8. References

Ashley, Kevin, Artificial Intelligence and Legal Analytics, Cambridge University Press, Cambridge 2017.

Bing, Jon, Legal Information Retrieval Systems: The Need for and the Design of Extremely Simple Retrieval Strategies, Computer/Law Journal, 1978, Issue 1, p. 379–400.

Blair, David C./Maron, M. E., An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System, Communications of the ACM, 1985, Issue 3, p. 289–299.

Dabney, Daniel, A Reply to West PublishingCompany and Mead Data Central on the Curse of Thamus, Law Library Journal, 1986, Issue 2, p. 349–350.

Dabney, Daniel, The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval, Law Library Journal, 1986, Issue 1, p. 5–40.

Gerson, Kevin, Evaluating Legal Information Retrieval Systems: How Do the Ranked-Retrieval Methods of WESTLAW and NEXIS Measure Up?, Legal Reference Services Quarterly, 1999, Issue 4, p. 53–67.

Grabmair, Matthias/Ashley, Kevin/Chen, Ran/Sureshkumar, Preethi/Wang, Chen/Nyberg, Eric/Walker Vern R, Introducing LUIMA: an Experiment in Legal Conceptual Retrieval of Vaccine Injury Decisions Using a UIMA Type System and Tools, Proceedings of ICAIL 2015, p. 69–78.

Hafner, Carole D., Conceptual Organization of Case Law Knowledge Bases, Proceedings of ICAIL 1987, p. 35–42.

Knapp, Melanie/Willey, Rob, Comparison of Research Speed and Accuracy Using Westlaw Classic and WestlawNext, legal Reference Services Quarterly, 2013, Issue 1–2, p. 126–141.

Mart, Susan Nevelow, The Case for Curation: The Relevance of Digest and Citator Results in Westlaw and lexis, Legal Reference Services Quarterly, 2013, Issue 1–2, p. 13–53.

Maxwell, Tamsin K./Schafer, Burkhard, Concept and Context in Legal Information Retrieval, Proceedings of Jurix 2008, p. 63–72.

Sormunen, Eero, Extensions to the STAIRS Study – Empirical Evidence for the Hypothesised Ineffectiveness of Boolean Queries in Large Full-Text Databases, Information Retrieval, 2001, Issue 3–4, p. 257–273.

Turtle, Howard, Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance, Proceedings of SIGIR 1994, p. 212–220.

van Noortwijk, Kees, Integrated Legal Information Retrieval: New Developments and Educational Challenges, European Journal of Law and Technology, 2017, Issue 1, p. 1–18.

van Opijnen, Marc/Santos, Cristiana, On the Concept of Relevance in Legal Information Retrieval, Artificial Intelligence and Law, 2017, Issue 1, p. 65–87.