Information Retrieval for Evidence-Based Policy Making applied to Lifelong Learning

Jérémie Clos(✉), Rong Qu, and Jason Atkin

Computational Optimisation and Learning Lab,
University of Nottingham
{jeremie.clos, rong.qu, jason.atkin}@nottingham.ac.uk

Abstract. Policy making involves an extensive research phase during which existing policies which are similar to the one under development need to be retrieved and analysed. This phase is time-consuming for the following reasons: (i) there is no unified format for policy documents; (ii) there is no unified repository of policies; and (iii) there is no retrieval system designed for querying any repositories which may exist. This creates an information overload problem for policy makers who need to be aware of other policy documents in order to inform their own. The goal of this work is to introduce a novel application area for studying information retrieval models: the information seeking phase of policy design, applied to life-long learning policy-making. In this paper, we address this problem by developing a common representation for policy documents, informed by domain experts, in order to facilitate their indexing and retrieval by users. This position paper highlights the research questions that we aim to answer in our future work and the dataset that we intend to use to do so. Our main contribution is the creation of a unified dataset of policy interventions which can be used for highly specialised information retrieval tasks, and will be released in order to provide the field with the first unified repository of policy interventions in adult education.

Keywords: Domain-specific search · Information retrieval · Case-based reasoning.

1 Introduction

Evidence-based policy making requires the design of policies based on not only ethical and practical goals, but also on evidence in the form of past attempts and their measured results in order to achieve two objectives: to design policies that are the most likely to have the desired effect, and to anticipate side effects that were observed on previous attempts and that might not have been considered.

However, policy making using the context of bounded agency shifts the focus from directly implementing the effects that we want to observe, to instead identifying the barriers and limitations that prevent the desired effects from happening, and implementing changes that diminish the power of such barriers or encourage desired behaviour. This complex interplay of social factors makes it difficult to compare policies, since multiple policies with the same final intent

(e.g., improving computer literacy in a specific target age range) might use different strategies to arrive at the same desired final result. This creates a particular difficulty for a user attempting to search for policies similar to the one they are trying to design: similarity is not defined with respect to content alone, but with respect to a set of conceptual dimensions which characterise the field of policy making.

In this work, we propose an approach to solving this problem using a high-level representation and a similarity function designed by domain experts for the explicit purpose of storing and retrieving policy documents. We use this approach to point to future research directions, and we discuss the settings of our upcoming evaluation.

2 The Policy Data Model

Our representation scheme was designed in two steps: (1) elicitation of low-level attributes by domain experts in lifelong learning policies, and (2) design of a lower dimensional feature space representing the low-level dimensions while reducing its sparsity. Our objective is to perform the matching phase on the four high-level attributes, while keeping the low-level features for the purpose of presenting information to the user. In this section we first mention the low level features and their elicitation, before moving on to the high level features and how they group low-level features into natural high-level concepts.

2.1 Descriptive Feature Model

A team of domain experts defined 78 attributes to describe the context of a policy. Those attributes represent multiple aspects of a policy, the most important of which are its geographical constraints (e.g. geographical code of the location, rurality/urbanity of the intervention), the socio-economic status of its participants (e.g. social status, social class, employment status), and salient features of the intervention itself (e.g. size and duration of the intervention, funding available). Focusing on those aspects allows the policy searcher to contrast multiple results and manually weigh which attributes might be more relevant to their own policy, which would not be possible if comparing large blobs of text from the source documents.

2.2 Policy Retrieval Model

The numerous descriptive features are then grouped into a higher-level policy model for the purpose of retrieval. A grand total of four features were judged to be sufficient to describe policies in a reasonable and retrievable way: target groups, aims of the intervention, activities performed during the intervention, and location of the intervention. A policy can possess more than one of each attribute (e.g. multiple target groups, or multiple activities).

Target Groups A target group represents a specific characteristic of the demographics targeted by a potential policy. It can take a specific set of values such as Gender, Ethnicity, Disability status, Age range, or more.

Aims The aim represents the explicit goal of the intervention described in the document. Such goal is related to a specific barrier between the target groups and the labour market that can be reduced through the activities of the intervention. A policy searcher who is proposing a new policy aiming to provide experience might want to find other approaches in culturally similar locations that aimed to reduce the same barrier in order to contrast with their own proposal.

Activities The activity focuses on how the aim is achieved, i.e. the activities that were performed during the intervention.

Location Location represents the geographical boundaries of the intervention described in the document. While the low-level feature model described in the previous section possesses a regional geographical code, the policy model uses a country-based geographical code. The reason for this is that the legal context does not change enough from region to region to justify the differentiation.

3 Query Models for Policy Search

Establishing the form of a query for policy searchers first comes from defining the typical profile of a searcher. We identify two phases in the process of evidence-based policy making, which parallel activities in Ellis' model of information seeking [3, 4] and phases of Kuhlthau's information search process model [6].

1. **Exploration** corresponds to the exploration phase in Kuhlthau's information search process, and to the browsing activity in Ellis' model of information seeking. The searcher has a vague idea of their aim, and they seek to explore the policies that have any degree of similarity with that aim, in order to refine their explicit information need ;
2. **Exploitation (comparing and contrasting)** is closer to the information collection step of Kuhlthau's information search process: the searcher has refined their information need, and formulates a more precise query. They are essentially filtering policies based on specific contextual constraints (geographical situation, socio-economic status), in order to compare and contrast their major differences. They are interested in high precision results more than recall rate, since they have already formulated the core of their proposed policy. It maps closest to the Differentiating activity in Ellis' model of information seeking: the sources are identified, and the information seeker uses their knowledge to judge their relative relevance.

These information seeking activities lead us to define three approaches for querying the document base: (1) the **free-text query**, suited for the exploratory stage, allows the searcher to match their query to any attribute with no restriction ; (2) the **free-text structured query**, i.e. free-text search over attributes, where the query is divided in four different fields and the relevance function is composed of a linear combination of similarities over each field, weighted by a user-defined preference weight in order to let the searcher define the priority of each field ; and (3) the **constrained structured query** is a more restricted version of the structured query, where the possible values are restricted to the

existing ones in the database and a predefined similarity, designed by domain experts, is used for the retrieval phase. Similarly to the free-text structures queries, the relevance function is computed as a linear combination of the similarities of each attribute, using user-defined weights. In this section we go over those three activities and the querying models and relevance functions associated to them.

3.1 Free-Text Queries

Free-text querying aims to completely focus on exploratory searches. The relevance function is defined in (1), where f_i refers to attribute i, d_i refers to the field i of the current document, q refers to the query document, and sim refers to a cosine similarity with tf-idf weighting [7], a standard information retrieval baseline. Simply explained, the relevance is the maximum similarity between the entire query and each of the four fields of the documents.

$$rel_{ftq}(d, q) = \max(\text{sim}(d_1, q), \text{sim}(d_2, q), \text{sim}(d_3, q), \text{sim}(d_4, q)) \tag{1}$$

3.2 Free-Text Structured Queries

Free-text structured querying serves as an intermediate step between fully structured and constrained queries and completely unstructured free text queries. It possesses the advantage of free-text query in that it is a high recall approach, but with the extra restriction of forcing the user to decompose their query into multiple fields. It is defined in (2).

$$rel_{fsq}(d, q) = \sum_{i=1}^{|f|} \beta_i \times \text{sim}_{f_i}(d, q) \tag{2}$$

Simply put, the relevance is calculated by a linear combination of similarity between each field f_i of the query q and each field of the corresponding document d, weighted by a preference weight set up by the user β_i.

3.3 Constrained Structured Queries

Constrained structured queries trade the free text fields for a list of choices extracted from existing cases in the database. This lets us use manually designed similarity tables that encode expert knowledge. The relevance score is described in (3), where exp-sim corresponds to the expert-designed similarity. This similarity is passed to a max operator due to the fact that a case might have multiple entries for a given field, and as such the relevance score needs to take the maximum value observed among all those entries.

$$rel_{csq}(d, q) = \sum_{i=1}^{|f|} \beta_i \times \max(\text{exp-sim}_{f_i}(d, q)) \tag{3}$$

The relevance is calculated as a linear combination of those expert-designed similarities for each feature, weighted by a user-defined importance weight β.

4 Evaluation

In this section we discuss the research questions we seek to answer.

RQ1 *Does enforcing structure in the query help match more relevant documents for policy searchers?* This research question can be answered by comparing the relative performance of the free-text structured queries against the free-text queries, given identical query content. An expected result if the structure improves the quality of the search is that the free-text structured queries would outperform complete free-text queries.

RQ2 *Does an expert-designed matching function perform better than a traditional statistical matching function on standard information-seeking tasks for evidence-based policy making?* This research question can be answered by observing the relative performance of the free-text structured queries against the constrained structured queries, given identical queries. An expected result, if the expert-designed matching function is well-suited to the task, is that the constrained structured queries would outperform the free-text alternative which is based on statistical knowledge.

5 The ENLIVEN Dataset

Our evaluation will use a new corpus, the ENLIVEN dataset, composed of 224 cases assembled from previous works in the field of policy making for lifelong learning as part of the ENLIVEN Horizon 2020 European project[1]. They were analysed by a team of domain experts and represented in the proposed high dimensional feature space, before being reduced to the higher conceptual one.

6 Background and Related Works

Structured information retrieval (SIR) focuses on the retrieval of information from structured and semi-structured document bases, such as XML documents [8]. SIR queries can contain structural information, which provide a matching criterion in itself. Case-based reasoning (CBR) is an artificial intelligence methodology that focuses on solving problems by retrieving similar problems with an existing working solution, and adapting them to fit the new problem [2, 5].

Structured Information Retrieval Traditional question answering methods typically retrieve a large number of documents using a bag-of-words model before doing post processing, which creates a computational bottleneck. Structured question answering solves that bottleneck by pre-processing the documents with multiple annotators such as a semantic parser and named entity recogniser [1]. The querying system then not only matches up the query to the answer, but also filters for annotation structures that denote a relevant information need. Similarly in our case, a free-text query would need to be categorised to predict the type of information need that it refers to, and then be matched against the corresponding attribute(s) of the document base.

[1] https://h2020enliven.org/

Case-Based Reasoning In the context of case-based reasoning (CBR), textual case-based reasoning comes closest to our work. Textual CBR (TCBR) focuses on the application of the CBR methodology on textual case bases [9]. The problem matching is done by comparing text-derived features in order to retrieve a typically textual solution. TCBR differs from information retrieval in goal and context only, and much of the techniques developed for information retrieval systems are used in TCBR systems.

7 Conclusion and Research Directions

In this work we introduced the problem of information retrieval in the perspective of a novel application: the design and development of socio-economic policies. We discussed a policy data model designed by a committee of experts in order to effectively and efficiently retrieve such policies during the information seeking stage of policy design. Finally, we turned our attention to the research questions that we will answer in future works, the corpus that we collected in order to do so, and briefly discussed the background area of our research.

Acknowledgements This work is funded by European Union H2020-YOUNG-SOCIETY-2015 (Grant agreement no. 693989). We would like to acknowledge and thank the people involved in the collection and categorisation of the ENLIVEN dataset.

References

1. Bilotti, M.W., Ogilvie, P., Callan, J., Nyberg, E.: Structured retrieval for question answering. In: Proceedings of the 30th annual international ACM SIGIR conference. pp. 351–358. ACM (2007). https://doi.org/10.1145/1277741.1277802
2. Craw, S.: Case-Based Reasoning, pp. 180–188. Springer US, Boston, MA (2017). https://doi.org/10.1007/978-1-4899-7687-1_34
3. Ellis, D., Cox, D., Hall, K.: A comparison of the information seeking patterns of researchers in the physical and social sciences. Journal of documentation **49**(4), 356–369 (1993). https://doi.org/10.1108/eb026919
4. Ellis, D., Haugan, M.: Modelling the information seeking patterns of engineers and research scientists in an industrial environment. Journal of documentation **53**(4), 384–403 (1997). https://doi.org/10.1108/eum0000000007204
5. Kolodner, J.: Case-based reasoning. Morgan Kaufmann (2014). https://doi.org/10.1016/c2009-0-27670-7
6. Kuhlthau, C.C.: Seeking meaning: A process approach to library and information services. Libraries Unltd Incorporated (2004)
7. Salton, G., McGill, M.J.: Introduction to modern information retrieval. mcgraw-hill (1983)
8. Schütze, H., Manning, C.D., Raghavan, P.: Introduction to information retrieval, vol. 39. Cambridge University Press (2008)
9. Weber, R.O., Ashley, K.D., Brüninghaus, S.: Textual case-based reasoning. The Knowledge Engineering Review **20**(3), 255–260 (2005). https://doi.org/10.1017/S0269888906000713