A CLOUD-BASED CONTEXT AWARE INFORMATION RETRIEVAL INFRASTRUCTURE FOR MANAGEMENT OF DISTRIBUTED LOCAL AUTHORITY DATA

Lewis Mc Gibbney , Post-Doctoral Research Scholar, lewis2@stamford.edu
Stanford University, State of California, USA

Bimal Kumar, Professor, B.Kumar@gcu.ac.uk
School of Engineering and Built Environment, Glasgow Caledonian University, Glasgow ,UK

ABSTRACT

In recent years the issues surrounding sustainable development of our land and resources has become (amongst others) a huge area of economic as well as historic interest. The Scottish planning system is designed to control positive change within our towns and cities making sure that the land is developed to every-one's long term interest. The notion of local planning allows more local ownership and decision making about the specific issues within a smaller area; however it creates a barrier for the cross organizational sharing and exchange of data which could be used to more effectively coordinate decisions made during the planning approval/review process. Additionally, a long standing argument has questioned inconsistencies in decision making which can be attributed to natural errors in human judgement. This highlights that the consistency of decision making needs to improve during checking and approval of planning applications. In this paper, we propose a cloud-based service that enables building control officers to obtain similar relevant planning applications (jobs) based on document parameters e.g. location. This enables officers to retrieve and consult similar planning proposals ranked on geographic location as one factor in an attempt to improve judgments leading to higher quality decision making. The novelty of such a system is that it utilizes an implementation of the canopy clustering algorithm to rank relevant documents. Such a system could be extended to rank relevant documentations on duration, use of building materials, resourcing and so forth.

Keywords: building design, planning, design checking, data retrieval, cloud-based service

1. INTRODUCTION

The aim of the introduction is to create an argument for improved integration of local authority data with a specific focus on building and planning applications. Justification can be attributed to the inherent inconsistencies which are present as a result of a failure for the current system to acknowledge the degree of human inconsistency which occurs within the process of reviewing such applications. This cannot be solely attributed to human error at building control officer level, however it should be noted that a lack of information sharing between local authorities leads to the existence of information siloes which in turn increase the likelihood of inaccurate, incorrect, or inconsistent ad-hoc decision making at building control officer level. One extremely important aspect and a key part of the formulation of this argument can be attributed to the fact that 36 local authorities in Scotland have to all comply with one set of regulations, however the regulations do not consider geographically distributed, historically and traditionally rendered building practices. Such practices are on occasion unique to one local authority or area (such as use of certain materials for aesthetical purposes), can be limited to a subset of localities or can indeed be applicable (such as fire regulations) across the entire country. A primary example is that real estate within the centre of Glasgow City has historically been built in Sand Stone of varying degrees (usually red or blonde); however Aberdeen in conflict has inherited the same nature of construction with preference being sided towards Granite Block due to local availability of such a resource. The regulations do not provide guidance for such discrepancies, therefore it is up to the building control officer to review the plans and specifications (regardless of

how detailed in nature) submitted as part of the building or planning warrant application and make a decision based on the knowledge within the immediate environment e.g. the office she works within. This knowledge is typically scattered around in several locations and can be one of several data types as illustrated in the next section.

2. BACKGROUND ISSUES

It is well known that the amount of information the building control officers have to consult in the process of checking a submission before issuing the building warrant is huge. The following diagrams (Toshner, 2010) are mind mapping diagram snapshots indicating the hugely complex and voluminous information to be referred in processing a typical building warrant application. These mind maps are very kindly provided by South Ayrshire Council in Scotland. Of importance is Figure 2 which displays documents from within the Building (Scotland) Act category. The icons associated with each document represent where it is located, namely; documents associated with an Internet Explorer icon represent out links which are available on the Web, and Scottish Government (SG) links represent documents directly linked to the SG web site.

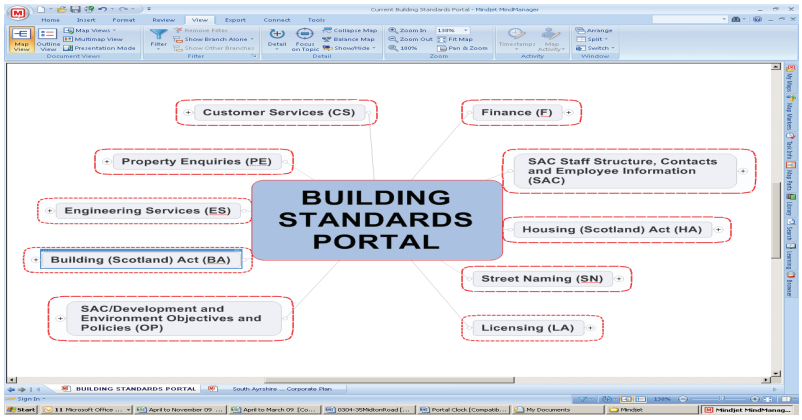


Figure 1: An example Building Standards Portal: Documents pertaining to heterogeneous and sometimes disconnected domains are grouped under topic category.

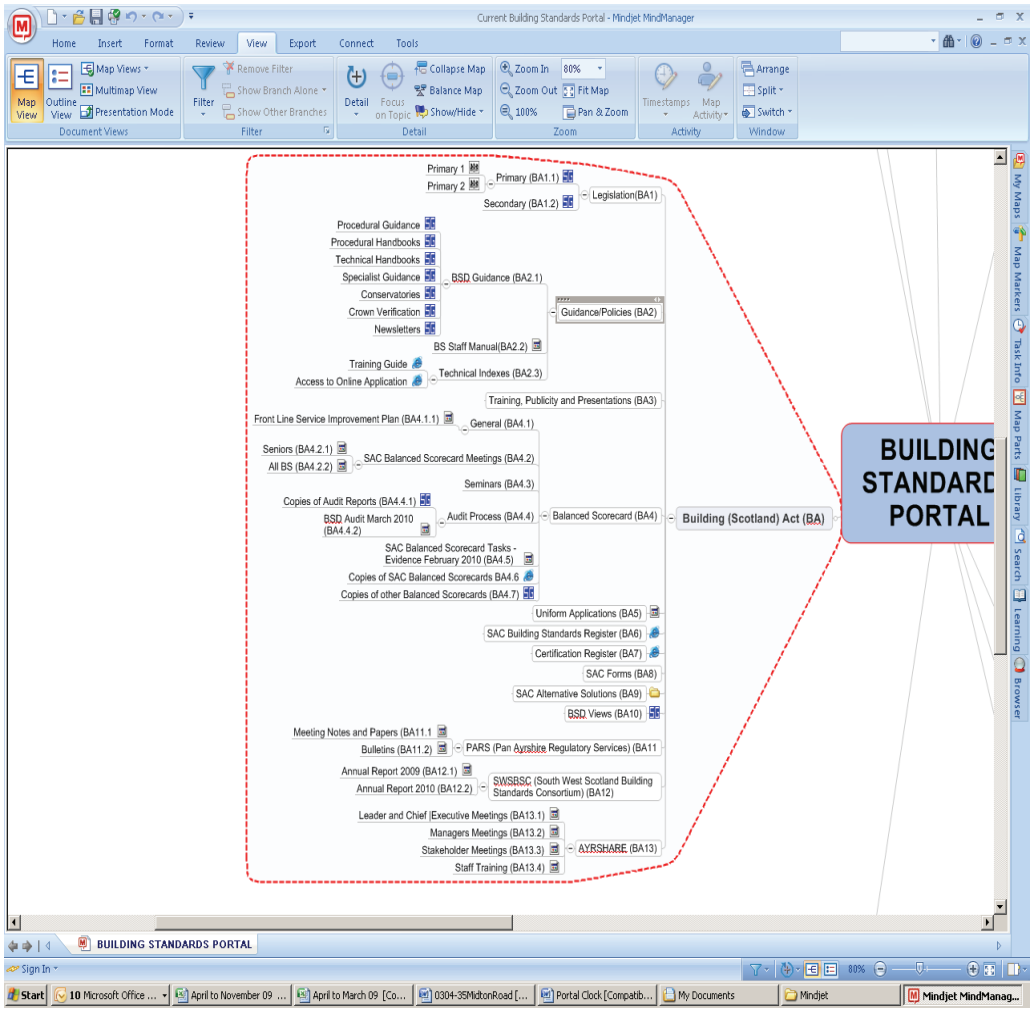


Figure 2: Tree structure representing available documents grouped under the Building (Scotland) Act category.

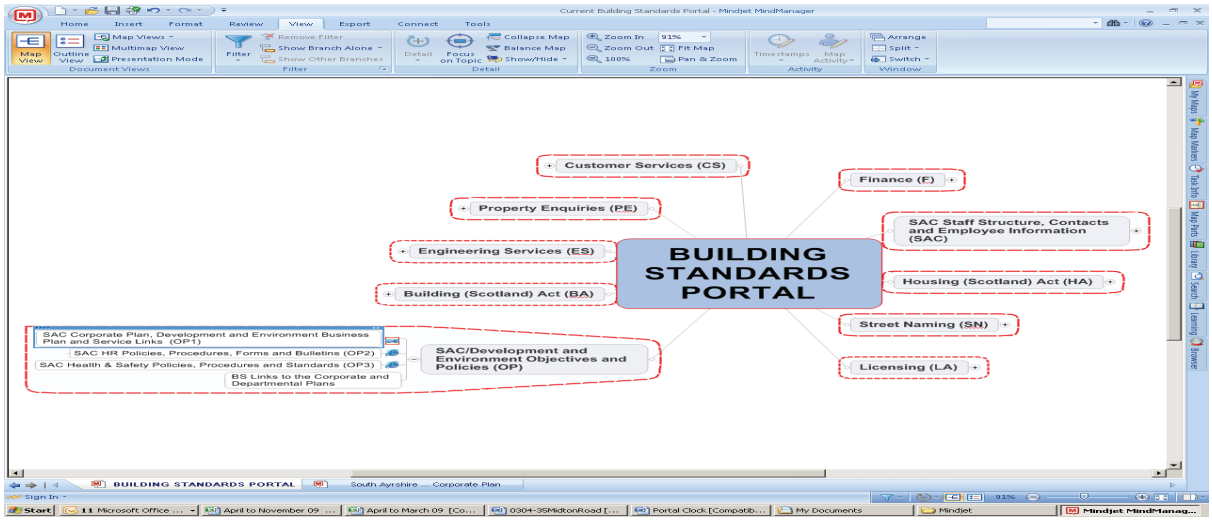


Figure 3: Branch expansion of the South Ayrshire Council/Development and Environment Objectives and Policies category.

Figure 3 represents quite clearly the scale of tasks and responsibilities of Building Control departments across Scotland, where officers are tasked with responsibilities well outside of the AEC domain. Such domains include environmental development, finance, community, private and social housing, customer services, street naming, licensing, etc. Figure 4 is a screenshot of a development agenda which concerns planning and local services.

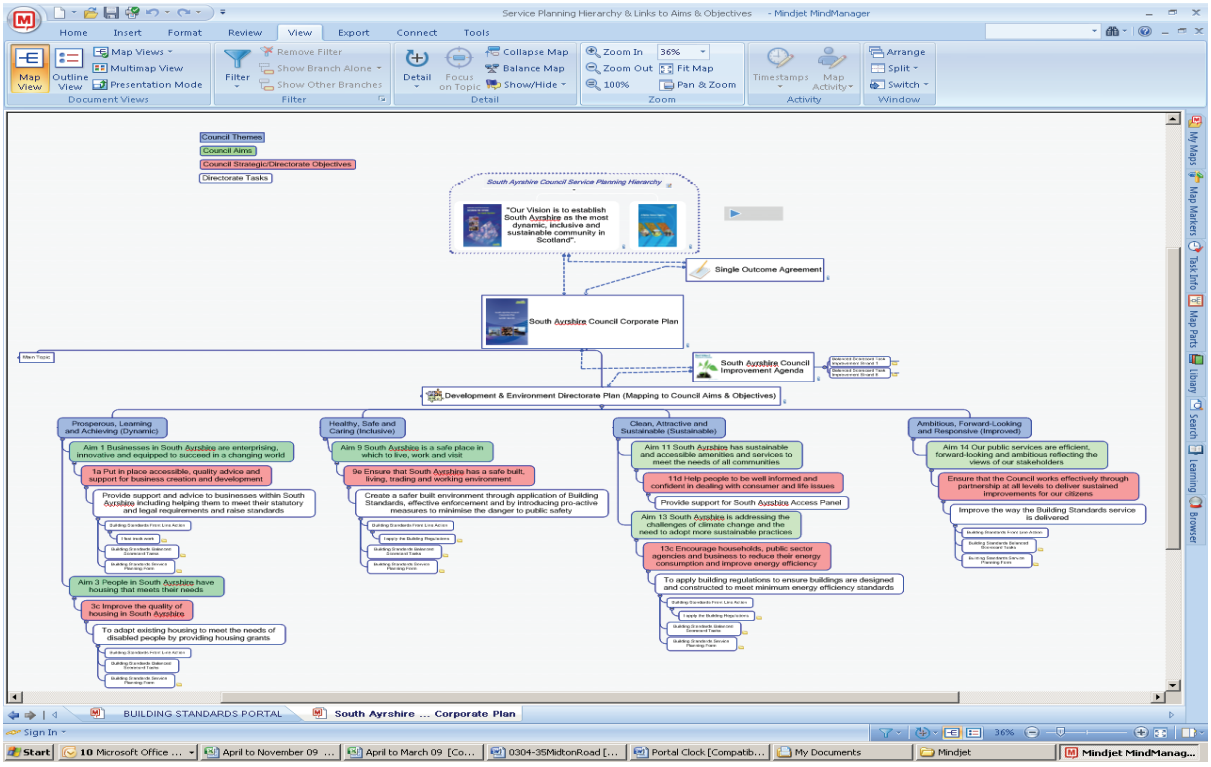


Figure 4: Example development agenda.

It is, therefore, clear that an effective document and content search engine will serve the building warrant process enormously. As mentioned earlier in section 1, context-dependent information like location, weather etc. is not fully covered by the Building Regulations (Building (Scotland) Regulations, 2004) and hence the reason for referring to other sources of information by the officers. At the moment, this process for searching for the appropriate documents and relevant information within those documents is dependent on one's experience and knowledge of the domain. This frequently causes issues particularly for the relatively new personnel taking on these tasks. The authors have developed a comprehensive framework for authoring (and therefore searching) regulatory documents and a detailed description of its implementation can be found elsewhere (McGibbney and Kumar, 2013). In the following section, we discuss an application of the canopy clustering algorithm for searching large multi-dimensional datasets, which addresses the problem discussed above.

3. UTILIZING GEOGRAPHICALLY DISPERSED PLANNING DATA AS A DRIVER FOR IMPROVED DECISION MAKING

When a planning permission/building warrant application is submitted, the physical application consists of several individual but essentially related documents. We are, of course, referring to builders/engineers specification, architects drawings, photographs, written commentary on certain technical aspects of the application which may or may not cause a discrepancy within the decision making process, and any other accompanying documentation which an applicant should see fit to include. When one considers that ad-hoc decision making regarding 'new applications' could

potentially benefit from retrieving such resourceful information and subsequently the addition information added to the application by building control professionals past and present, a more appealing argument results from enabling access to such information.

The system currently in use is measured on the time it takes for applications to make their way through and for decisions (relating to eventual verification or denial) to be made. The system is not evaluated on the basis of good decision making. Systems such as the one we are proposing essentially aim to re-balance that scale.

The proposed approach will facilitate 'similar' applications being made available to building control officers. These can then be searched on various criteria as appropriate in case to case. For example, one of those key criteria could be the 'location' of the building as it is an important feature to be considered. This is because as well as making available similar jobs (applications), we want to refine the granularity of the decision also based on ranking similar jobs by the geographically closest ones. This is to say that (for example) the decision to grant a planning application in Town A for the demolition of an external dwelling house (outhouse) to make way for a new extension structure to the main property and the addition of dormer windows and solar panels to this new structure, would most likely benefit most from information concerning similar applications (and therefore decisions) which were made in other towns closest to Town A as opposed to ones farther away. The justification is that parts of the country within close vicinity to each other tend to look more alike and suffer from the same/similar issues than those further afield. This is historic and inherited. It is the job of the building regulation and planning system to protect this as a matter of public and historic interest. It should be mentioned that the core ideas driving our approach is similar to case-based reasoning approaches (Kolodner,1993) used in several domains for aiding and improving decision making processes (Raphael and Kumar, 2001).

Based on the above discussion, the problem being addressed by our approach and the proposed solution is summarized below.

Problem:

- Data is located in numerous *in-house and external* data sources and access to the data is restricted.
- The data resides in several (most likely relational) databases. The planning data between offices will most likely contain similar semantics but will most likely not conform to the same structure.
- What will it take to develop a test corpus for use in the framework?

Solution:

- Load all (or as much of the data) into the Hadoop File System (HDFS); HDFS is a distributed file system that provides high-throughput access to application data for use within the Hadoop framework. The benefit of this is that we can give the data a uniform representation but retain its data semantics which are key to enabling the Canopy Clustering implementation.
- Use the processing model provided by Hadoop's MapReduce (MR) implementation to execute distributed processing of the (large) data set residing in HDFS.
- Utilize an implementation of the Canopy Clustering algorithm which uses the MR paradigm for data processing. It should be noted that although running time of this particular approach is not one of our primary concerns, as we are using MR to parallelize computation the processing improves computation over large datasets which is typically suited to local authority datasets.

4. A CANOPY CLUSTERING APPROACH FOR PLANNING PROPOSAL DOCUMENT ASSOCIATION

Canopy Clustering [McCallum et al., 2000] is a relatively simple, fast and surprisingly accurate method for grouping objects into clusters; in this case objects can be considered as planning or building warrant applications. All objects are represented as a point in a multidimensional feature space. The algorithm uses a fast approximate distance metric and two distance thresholds T1 > T2 for processing. The basic algorithm is to begin with a set of points and remove one at random. Create a Canopy containing this point and iterate through the remainder of the point set. At each point, if its

distance from the first point is < T1, then add the point to the cluster. If, in addition, the distance is < T2, then remove the point from the set. This way points that are very close to the original will avoid all further processing. The algorithm loops until the initial set is empty, accumulating a set of 'canopies', each containing one or more points. A given point may occur in more than one Canopy. Canopy Clustering is often used as an initial step in more rigorous clustering techniques, such as K-Means Clustering (MacQueen, 1967). By starting with an initial clustering the number of more expensive distance measurements can be significantly reduced by ignoring points outside of the initial canopies.

The key idea of the canopy algorithm is that one can greatly reduce the number of distance computations required for clustering by first cheaply partitioning the data into overlapping subsets, and then only measuring distances among pairs of data points that belong to a common subset (McCallum et al., 2000).

For the clustering implementation we utilize Apache Mahout; a scalable machine learning library for clustering, classification and batch based collaborative filtering implemented on top of Apache Hadoop using the MapReduce paradigm. Some specifics of the implementation are described in the following sections however first lets discuss a high level overview of how the Canopy Clustering algorithm is implemented within Mahout. Effectively there are two stages, namely:

- Canopy generation: effectively the process of identifying "a subset of the elements (i.e. data points or items) that, according to the approximate similarity measure, are within some distance threshold from a central point. Significantly, an element may appear under more than one canopy, and every element must appear in at least one canopy." (McCallum et al., 2000), and
- Clustering: assigning a weight and a vector to each data point within the set. When taken together, they carry the probability that each data point is a member of the given canopy

5. ADAPTING CANOPY CLUSTERING TO THE LOCAL AUTHORITY DOMAIN

It is an unfortunate reality that access to local authority departmental databases is restricted for research purposes. To overcome this hurdle, we therefore replace the representation of real building warrant applications stored in database with webpages from each local authority website. We can easily construct a HDFS compatible web graph of target webpages using Apache Nutch[1] a a highly extensible and scalable open source web crawler software project. When thinking about how we wish to process these documents, in terms of their document structure, both entities are actually not too dissimilar. If we consider that each document can be located by a key and that this key will be the document URI[2], then for database entries let database keys map to webpage URL's. This way we can identify individual documents based on their URL values as they are unique within the overall collection. Additionally, each document has fields representing title, content, content type, text, in links, out links, metadata, etc. In reality building warrant applications express similar document semantics which we can utilize within this study. We also however, add one additional metadata entry to each webpage.

```
fetching: http://www.eplanning.north-ayrshire.gov.uk/OnlinePlanning/applicationDetails.do?activeTab=summary&keyVal=M
parsing: http://www.eplanning.north-ayrshire.gov.uk/OnlinePlanning/applicationDetails.do?activeTab=summary&keyVal=MQ
contentType: application/xhtml+xml
content :       13/00454/PP | Erection of attached garage to side of detached dwelling house | Site To South East Of
title : 13/00454/PP | Erection of attached garage to side of detached dwelling house | Site To South East Of
host :  www.eplanning.north-ayrshire.gov.uk
tstamp :        Sat Aug 03 19:14:23 PDT 2013
canopyGroup :   NorthAyrhsire
url :   http://www.eplanning.north-ayrshire.gov.uk/OnlinePlanning/applicationDetails.do?activeTab=summary&ke
```

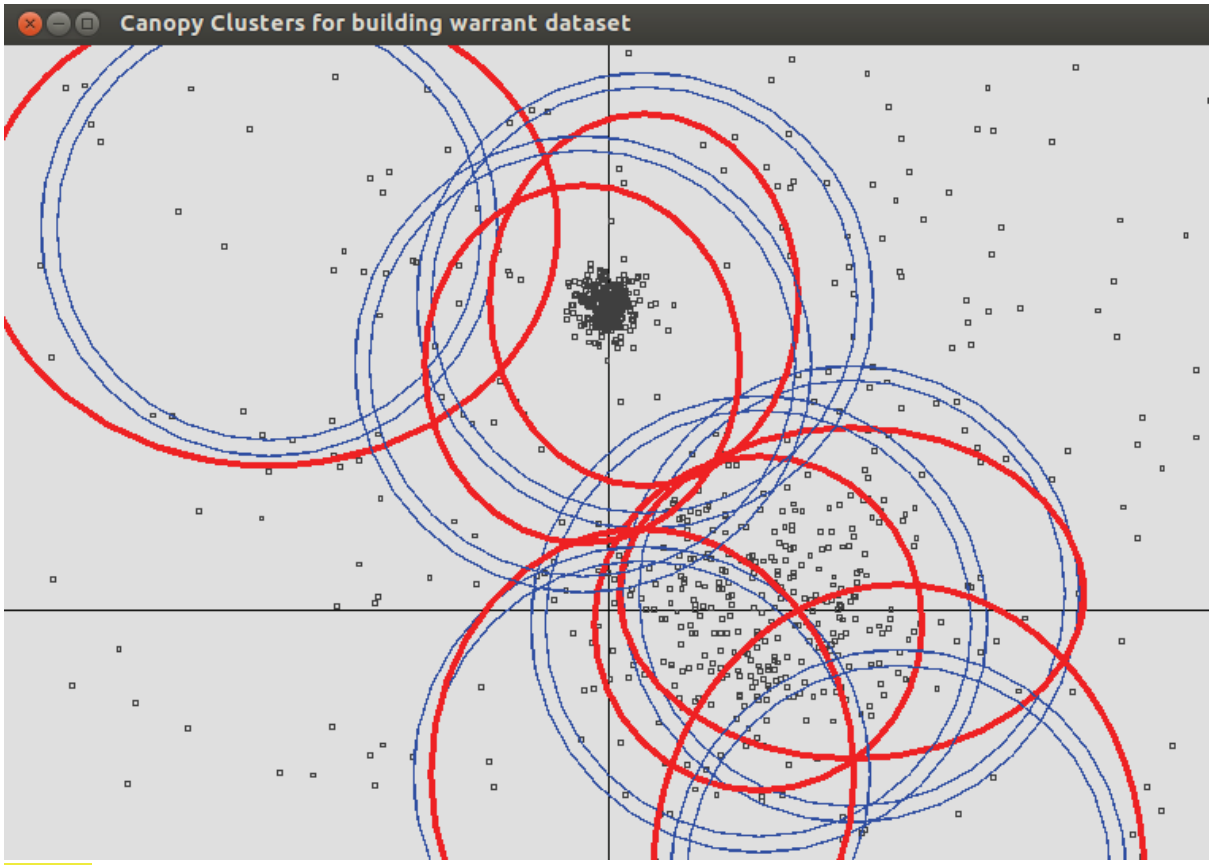
Figure 5: A snippet of crawl data obtained from particular building warrant application. The *canopyGroup* field and the corresponding value can be seen

1 http://nutch.apache.org

2 Uniform Resource Identifier: http://www.w3.org/Addressing/URL/URI_Overview.html

This entry corresponds to the actual local authority which the web page belongs to (e.g. which website it was fetched from). When we run the Canopy Clustering implement the addition of this field is critical to ensuring that a cheap distance measure can be used to create initial overlapping subsets. As per (McCallum et al., 2000), the term canopy is derived from these subsets. For each web site domain (each domain representing a local authority) we ensure that a static field *canopyGroup* and a corresponding value is added to the fetched document. Figure 5 shows a snippet of crawl data from a warrant application filed in North Ayrshire. One can clearly see the static field *canopyGroup* and its associated value *NorthAyrshire,* relating to the local authority within which this particular building warrant application was filed.

When we apply this method of assigning cheap distance metrics to web pages within the data set we are able to satisfy the first stage of the overall clustering process. Some initial clusters, produced from a subset of our dataset can be seen in Figure 7. In this screen shot, we can see that in all seven canopies (represented in red) have been produced for the last iteration of clustering. The blue canopies represent previous iterations (prior to the last iteration) where data points have been removed and the next iteration executed. We are finding that, in addition to looking at the generated output, being able to visualize the canopies in this way is enabling us to understand how clusters can converge upon a solution over multiple iterations.



图名？

6. SUMMARY AND DISCUSSION

The current approaches to processing building warrant applications in most parts of the world are lengthy. This is mainly because of lack of automated or semi-automated processes for storing and retrieving relevant information from a vast array of documents to aid the decision making process. The most ‘advanced’ approach, in our experience, in Scotland was the use of a mind mapping software to model the structure of linked documents and sources of information without any search capability. An approach based on canopy clustering search algorithm has been proposed to address this problem and initial observations show that based on understanding data characteristics and

exposing such characteristics as similarity metrics, we are able to create canopies and cluster documents from a sub set of our data. Based on this these preliminary results, there is a stronger argument for cross-organizational information sharing concerning planning and building warrant applications within Scotland and further afield. The proposed approach is still under active validation and development and more detailed results will be available in due course.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support provided by Mr. M Toshner (Toshner, 2010) of South Ayrshire Council in Scotland. He very kindly provided the diagrams provided on pages 2, 3 and 4 which formed the basis for this work.

REFERENCES

Kolodner, J. Case-based Reasoning, Morgan Kaufmann Publishers, USA, 1993.

Mahout, https://cwiki.apache.org/MAHOUT/canopy-clustering.html

MacQueen, J. B. , "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07, 1967.

McCallum, A.; Nigam, K.; and Ungar L.H., "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178, 2000.

McGibbney, L. And Kumar, B., An Intelligent Authoring Model for Subsidiary Legislation and Regulatory Instrument Drafting within Construction and Engineering Industry, International Journal of Automation in Construction, Elsevier, Accepted for publication, 2013.

Raphael, B. and Kumar, B., Reconstructive Memory in Design Problem Solving, Research Monograph, Saxe-Coburg Publications, Edinburgh, ISBN No. 1-874672091, 2001.

Building (Scotland) Regulations, Scottish Building Standards Agency, 2004.

Toshner, M., Personal Communication, 2010.