

DISCO: Intelligent Help for Document Review

Jacki O'NEILL, Caroline PRIVAULT, Jean-Michel RENDERS, Victor CIRIZA,
Gregory BAUDUIN

Xerox Research Center Europe. 6 chemin de Maupertuis, 38240 Meylan, FRANCE
{Firstname.Lastname}@xrce.xerox.com

Abstract

This paper describes a tool for assisting lawyers and paralegal teams during document review in eDiscovery. The tool combines a machine learning technology (CategoriX) and advanced multi-touch interface capabilities to not only address the usual cost, time and accuracy issues in document review, but to also facilitate the work of the review teams by capitalizing on the intelligence of the reviewers and enabling collaborative work.

1. Introduction:

E-discovery is the forensic collection and production of all documents relevant to a legal case. In corporate litigation cases and government investigations, where the volume of documents which need to be collected, sorted and annotated is often huge, e-discovery can be an extremely costly and time consuming process. In this paper we present a prototype system, DISCO, which we have designed to facilitate the document review stage of e-discovery. Document review is the process of deciding which of the (often millions) of collected documents actually need to be produced (handed-over) to the opposing party. Document review can be considered to consist of a number of stages which can be organised in different ways (e.g. they may be done in sequential reviews or one after another in the same review). The stages typically comprise of 1) *responsiveness review* - deciding whether a document is pertinent (responsive) to the case or not (non-responsive) 2) *privilege review* - deciding whether a responsive document has content which should not be exposed to the other side. All responsive, non-privileged documents must be produced. 3) *Issues coding* - deciding which issues in the case a document pertains to. Although not strictly part of first level review, this may be carried out at the same time. Document review is a very costly process, typically involving large teams of lawyers and paralegals, often contractors, rapidly reading through and coding thousands of documents per day. Most of the documents they read are likely to be deadly dull, with only a tiny proportion of them responsive and even less of actual significance to the case. Thus the work is generally monotonous, and this when combined with the often poor working conditions of the review teams [1] means that the work is likely to be highly error prone [2]. However, people remain central to this work as there is not currently, nor in the foreseeable future, any technology which can *understand the semantic content of documents*. Unfortunately, the very reason that people are hired - their intelligence and ability to understand the semantic content of the documents and the contingencies of the legal case - is somewhat muted by the current conditions of work and technology they use.

Even given their limitations, document review is a fruitful area for technology development because the increasing amount of information being stored means ever increasing numbers of documents for review (e.g. a matter is typically 1-2 million documents, with many being over 6 million). The proliferation of language processing technologies - categorisers, clusterers, Natural Language Processing tools - would seem to be well suited to addressing this problem. Whilst being unlikely to fully automate review, technologies might be used to reduce amount of documents legal teams have to process, to speed up the processing or to increase the accuracy. Technologies should be chosen according to the contingencies of the case, e.g. if the risk of mistakenly producing a privileged document is not too high, bulk classification tools might

provide an acceptable balance of time, cost and risk; alternatively if required accuracy is high, then manual review is likely to remain the best strategy. Additional concerns for the legal profession when employing technologies include how understandable are their workings?, how accurate their output?, how explainable their results? and how much trust they themselves have in the produced output? Whilst it is known that manual review can be inaccurate [2] [13], it is considered as a known inaccuracy. Indeed, one of the questions that might arise with processing technologies, is even if both the technologies and the legal team would miss the same percentage of documents, one might expect the legal team not to miss any *key* documents because they are making judgements based on the meaning of the document in relation to the case – not according to some pattern of words therein. In actual fact, there has been little detailed study of review accuracy – either by technology or people– so a real comparison is hard to come by.

The prototype system, DISCO, described in this paper, aims to improve the accuracy and speed of manual review by capitalising on the intelligence of the reviewers whilst supporting their work with content analysis capabilities (clustering and categorisation) and new interface technology. Before we discuss our system further we will describe the current use of technology in document review.

2. Document review technologies

Many software applications have been designed for the litigation support market. Users can, for instance, sort the documents before starting the review or more advanced customization can be achieved through preliminary grouping (clustering) of the documents. The techniques for sorting and clustering are generally keyword based, which is powerful - especially for culling documents - but suffers from known limitations. The inaccuracy and limitation of keyword (including Boolean) search is due to numerous factors, such as lack of context-sensitivity, over- and under-inclusive results and the "knowledge acquisition bottleneck" i.e. the burden of manually writing the rules, and formulating and selecting keywords. In general if users have poorly defined or too complex goals, if they do not know what they are looking for, tools entirely based on keyword search do not provide adequate support. In the case of litigation, the lawyers determining the keywords upfront are not always familiar with the case terminology, but as the reviewers become familiar with the terminology, their lack of experience with the technicalities of keyword search tools mean they cannot easily review, change and extend the keywords, which runs counter to the assumptions behind the use of such tools. The volume of research in the IR field is considerable, but one can cite as a recent reference a real case discussion over the effectiveness of search terms in eDiscovery [9], the Sedona Conference's "Best Practices Commentary" on IR in E-Discovery [7], and [8] about the significance of affect in the interaction with IR systems. It is largely accepted that "Concept Search" can augment keyword searching by broadening a query to include synonyms, using a thesaurus to include terms with similar meanings or through linguistics. However it can increase over-inclusiveness and it does not alleviate the burden of the "knowledge acquisition bottleneck" mentioned above.

Sorting and clustering is also frequently based on meta-data information (such as date, e-mail address or document type) which is always useful and necessary but by essence does not rely on the textual content of the documents at hand. This suggests that keyword, concept and metadata search technologies are better used coupled with a global statistical approach and we propose here to use a machine learning approach.

It is also interesting to observe how current keyword search systems are used: it seems that in practice these tools tend to be used in a rather limited way. The usability of search engines is a research domain in itself [e.g. 10], but the technical background and culture of users does not always provide them with the necessary skills for using these kinds of systems to their maximum capabilities. Formulating a query can be difficult especially when Boolean logic is used. The legal community is now becoming more familiar with it, but making complex decisions of including

and excluding query terms to balance between recall and precision is not easy, especially for users unfamiliar with statistical notions. Contractors on document reviews particularly need guidance for using search tools properly [7]. In addition, the many possibilities of advanced query systems can increase the user interface complexity and have an impact on user acceptance. This suggests a need for intuitive and easy-to-use interfaces that conceal, in part, the complexity of the underlying technologies.

3. DISCO: Assisting reviewers through intelligent help

The considerations outlined above, that is a) the need for technological support to help with increasing volumes of documents, b) the strong requirements for both accuracy and accumulating trust in any document processing technology used and c) the need for intuitive interfaces enabling technology to be easily mastered and employed in the document review process – led to the design of DISCO our document review support system. It is a manual review system which was designed with 3 guiding principles:

- 1) Capitalise on the *intelligence of the document reviewers*. Currently manual review can be a soulless process, however peoples understanding of the semantics of the documents and the contingencies of the legal case are a vital component of accurate review. We aim to give reviewers better tools and to make the work more interesting, which we believe will improve accuracy and speed. The key player in the review task remains the human annotator who reviews each document.
- 2) Support the reviewers work with *machine learning technology*, rather than automating it. We use a clustering and categorisation technology to facilitate the organisation of the documents and enable learning from practice.
- 3) Capitalise on the *affordances of paper and electronic documents* to produce a natural, easy to use interface. Paper documents can be annotated, shuffled, put to one side, etc., electronic documents can be sorted, processed and searched. Given that lawyers are primarily reviewing documents and the desirability of having a natural, easy-to-use interface, we designed the interaction using the metaphor of paper documents and implemented the review assistant tool on a multi-touch surface device. This kind of touch screen not only provides natural interaction with the application through finger and hand contacts on the screen, but can contribute to hiding the complexity of the technologies in use. Moreover, it can enable collaborative reviews within a team.

With DISCO, we aim to improve both individual and collaborative review of documents to improve learning and accuracy. We will outline how we intend to do this in the sections below; first providing an overview of the CategoriX technology and how it will be used before describing the implementation and interface aspects. We finish with a discussion of this work.

3.1 CategoriX

Based on research from XRCE¹ and originally from PARC², CategoriX is a machine learning system that learns from human annotations to build probabilistic models. These are used to automatically infer the probable category(ies) of new incoming documents. (See [5] for an overview of machine learning in automated text categorization). CategoriX relies on a probabilistic generative model [4] which can be seen as a hierarchical extension to PLSA (Probabilistic Latent Semantic Analysis [3]) where both documents and words may belong to different categories. The basic assumption of PLSA is that there exists a set of hidden (“latent” or “unobserved”) factors which can underlie the co-occurrences among a set of observed words and

¹ Xerox European Research Centre

² Palo Alto Research Centre

documents. The aim is to find these parameters in order to build a *generative (predictive) model*. CategoriX (CX) follows a two-phase process: 1) the generative model is learnt from a subset of documents *manually* assigned to some categories in a pre-defined taxonomy. CX learns from a Subject Matter Expert (SME) decisions, in order to reproduce these annotations on new documents; 2) the generative model is used to infer the category assignment of new, unclassified incoming documents.

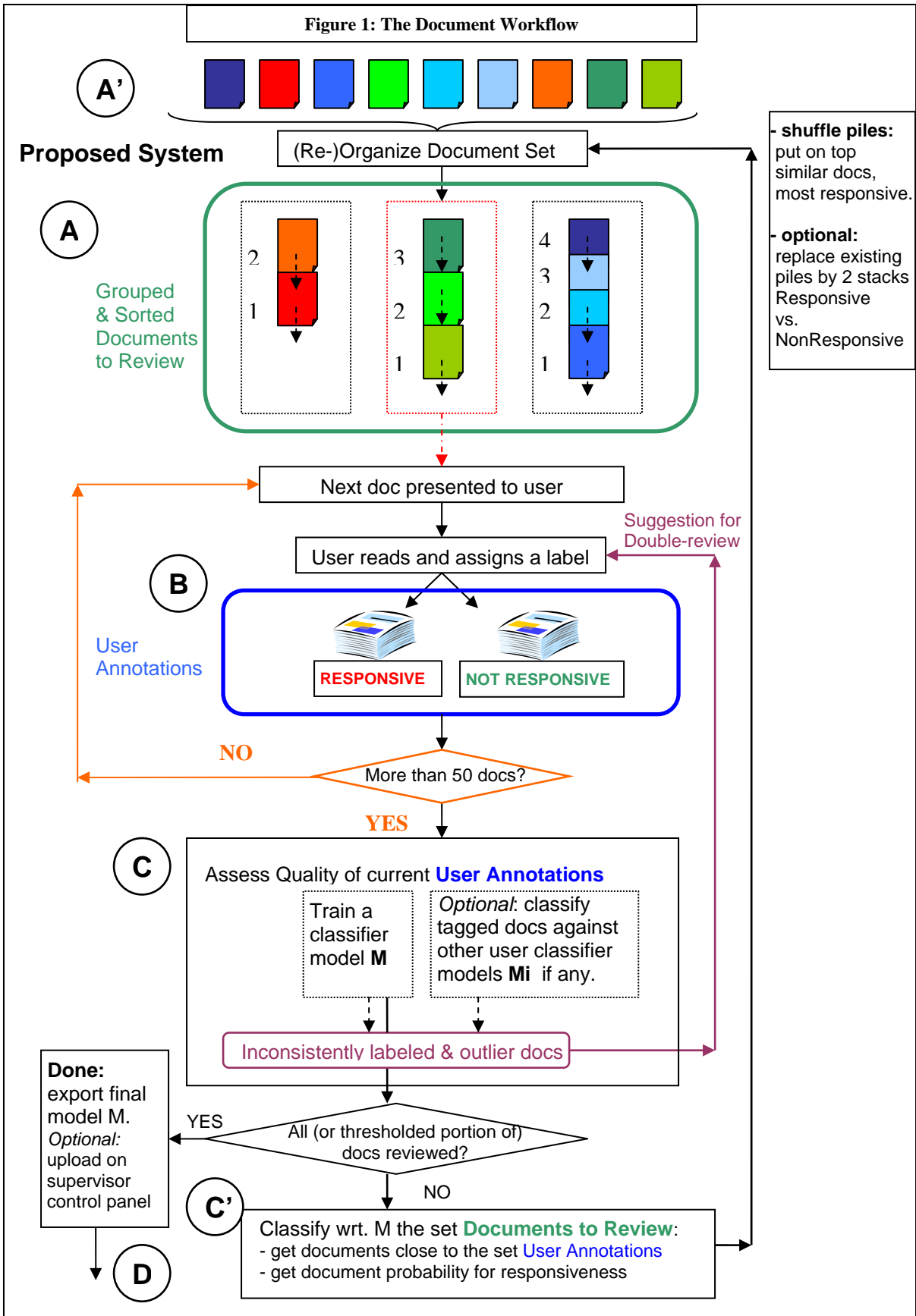
The primary output of CX when a document d is sent for evaluation, is a vector of probabilities against each model category: the value encountered at index i indicates the probability of category C_i given document d . This results in a “soft assignment” of the document to the model categories. Using simple or more advanced heuristic rules, a “hard assignment” of the document to one or more categories is easily derived, for instance assigning the document to the category with the highest computed probability.

As well as making use of the basic categorisation features, DISCO also uses advanced capabilities such as “outlier detection” and “mis-tagging detection”, plus document clustering. *Outlier detection* consists in detecting if a document is partially or strongly outside of the scope of a classifier model. This feature can flag up atypical or rare documents, as well as noisy elements. The *mistagging detection* can be used to analyze a set of SME’s annotations for consistency. When a document is flagged as “possibly mislabelled” by CX, a more appropriate category is also suggested. *Document clustering* is the process of automatically inferring groups of documents based on their content. In CX, clustering can be seen as the unsupervised counterpart of classification and is supported by the same underlying statistical model. Without supervision of an SME or prior annotation of a subset of data, documents are grouped in a user-defined number of clusters. As well as partitioning the documents into clusters, CX provides a generative statistical model of the set of clusters. This can be used as a classifier to route documents from a larger data set to the existing clusters, thus evading the problem of computational overhead when computing clusters for large data collections.

3.2 Using CategoriX for document review

Because CX bridges clustering and categorization through the same statistical representation, several scenarios can be derived for supporting document review. In this application, the key player in the review task remains the human annotator whose work is to read each document to decide if it is responsive, but the system is there to facilitate, speed-up and improve the review by: **(a) grouping documents** in a useful way - by topic or likelihood of responsiveness; **(b) sorting documents**, instead of presenting them in a random order; **(c) analyzing user annotations**, and providing feedback on tagging consistency; **(d) suggesting next documents to review**, e.g. similar to the ones already reviewed or inconsistently labelled; **(e) enabling tagging validation/cross-evaluation** (against other user annotations).

The system works as follows: A batch of documents is uploaded. To begin, the reviewer can create some groups either through automatic clustering or through automatic classification against a pre-existing model (if one exists). The user chooses which group of documents to start with, (e.g. based on group size or representative keywords (automatically determined by CX)), this opens a first document which the user reviews and assigns a label (e.g. Responsive or Non-Responsive (R/NR)). Once the user has reviewed documents up to a system parametrizable threshold (here defaulted to 50 docs) CX uses this as training material and automatically builds a statistical model of the current users annotations and returns some feedback: a subset of “possibly mistagged” documents as an evaluation of the consistency of the annotation, and atypical (outlier) documents. The user receives a suggestion to double-review these documents: she can either double check them or continue tagging new documents. When enough documents have been accumulated to train an accurate classifier model (a system parametrizable threshold defaulted



here to 100 docs), a binary categorizer R/NR is built and the user is given the possibility to automatically modify the grouping of the not-yet-reviewed documents into 2 batches: 1) “*possibly responsive*” and 2) “*possibly non-responsive*”. Every new user annotation falls into the document training set and enriches the automatic binary classification through an incremental retraining of the model. At the end, the global set of annotated documents is modelled as a classifier, assessed for consistency through “possible mistagging” and outlier detection, and exported for further re-use or cross-validation.

Figure 1 shows a representation of the document workflow in the system.

Stage A represents the un-organized set of documents that the user has to annotate. Documents are assigned to a reviewer by a review supervisor from a larger collection – either randomly, on the basis of metadata or through a preliminary clustering stage. This preliminary stage may be done in a number of runs to avoid problems of intractability. The collection is sampled, CX is applied to the sampled sub-part to group documents into N clusters. (e.g. for N reviewers). Then bulk-classification is applied to the rest of the collection, to route every document to a cluster. If a cluster is too large it can be split a posteriori. If the review supervisor has some prior knowledge of which documents are responsive, the clustering stage can be semi-supervised by pre-assigning some responsive documents of different types (e.g. issues related) to some empty clusters.

In **Stage A** the documents to be reviewed are 1) grouped: either by a) clustering, into an arbitrary number of clusters or b) classification against an existing model - in which case the grouping is made up of 2 piles “possibly R” and “possibly NR”, 2) ordered within each group, putting first the docs of most interest to the user. The organization in Stage A will be continuously modified, thanks to the statistical learning from the user’s tags as reviewing proceeds.

Stage B represents the set of documents tagged by the user up until now. User tagging is not restricted to binary classification; it can be of any kind, e.g. it can be the R/NR classification plus Privileged/Non-Privileged of documents tagged R, or it can be a multi-category classification scheme for “issues coding”, depending on the review organisation. Users can attach additional information to a document: highlighting text e.g. to show the motivation of the classification, adding a sticky tag for one or more pre-defined items such as ‘hot’, ‘key’, ‘confidential’, etc.

Stage C shows the statistical analysis of the user tagging from Stage B. It is used to assess the quality of the tagging and to provide feed-back for re-organizing Stage A. At Stage C, the classifier is trained from the tagging (Stage B) to create a categorizer model. This model is used to (a) re-sort groups in Stage A (by identifying not-yet-reviewed documents similar to those just tagged); (b) suggest re-review of documents in Stage B (by identifying possibly inconsistently labelled documents); (c) change the grouping of Stage A into two piles: “possibly R” vs. “possibly NR” documents when an accurate classifier model has been trained; (d) as part of multi-reviewer quality assessment in Stage D.

Stage D is the final modelling of all the user tagging. It can be exported for quality assessment by the review supervisor. The quality checking is twofold: self-assessment of the user model (e.g. list of possibly inconsistently labelled docs); cross-validation with other user models : model M1 of set S1 of docs issued from user#1 tagging, is applied to categorize set S2 of docs issued from user#2 tagging. Documents for which the tagging of S2 is inconsistent with the tagging predicted by M1 are flagged. In the same way docs S1 are processed by user model M2. In addition documents that are frequently inconsistently labelled among different reviewers are flagged as “difficult documents” and culled out for a double review. Documents that are flagged as “outliers” by each model can also be collected and culled out for a double review.

The above system can be used for responsiveness review, but it is naturally flexible enough to be used for privilege and issues coding. In addition the system helps review coordination, enabling better assignment of documents to reviewers (using clustering in Stage A’), and quality control

and results tracking since having a model trained from each reviewer's annotation provides some means to evaluate the global outcome of the review process. The percentage of possibly mislabelled documents can help rate a user's annotation result. In addition, cross-validation between the different reviewers can be obtained through stage D and the set of documents inconsistently labelled across the review team can be extracted and selected for re-review.

A litigation document workflow having similarities to DISCO was described in [6]. As with DISCO user annotations are accumulated during the document review before retraining a classifier. The application is however tightly linked to email processing, based on the extraction of metadata and features extracted through a simple string matching mechanism, based on user-defined keywords or expressions for the case at hand. Another online learning system can be found at [14] where the software continuously learns from the attorney judgments to enhance the accuracy of the automated classifiers. Without the technicalities of the proprietary machine-learning algorithm in use a real comparison is not possible. However, statistical clustering is not part of the workflow, and the "erroneous classification" seems to be achieved through measuring the disagreement between the system and human prediction while in CX mislabelling is achieved through a measure of entropy and level of ambiguity of the document between the different possible categories.

3.2 Natural interaction through innovative UIs

We chose to implement DISCO on an innovative multitouch device, as it seemed best suited to enabling an intuitive interface which hides the complexity of the technology, at the same time giving more freedom to the user to organize her working environment. Touch screen interfaces are becoming increasingly popular (iPhone, MS surface table, etc) as they provide a natural means of interaction. The display detects user contacts and translates their positions to associate them with the UI widgets and commands. We developed our system on the Multi-Touch G² - Touch Screen Technology from PQ Labs [11]. This device is made up of a 42 inch touch plate which is placed over a horizontal LCD screen making an interactive table.

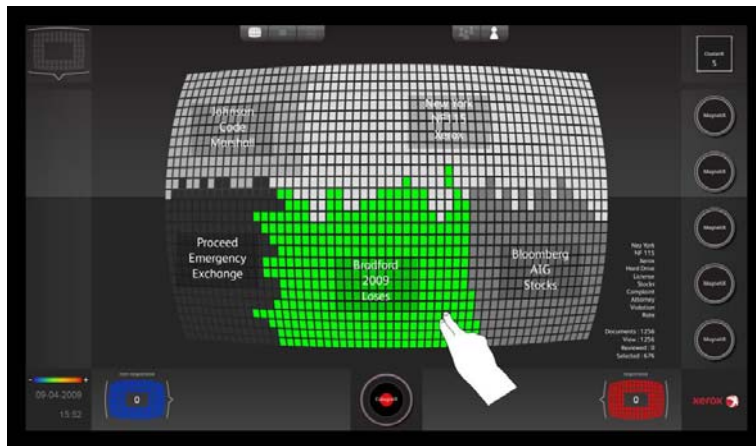


Figure 2: The wall view after ClusteriX applied

The user has two views onto the system; 1) the wall view presents the documents as a representative wall (Fig. 2) and is used to manipulate sets of review documents; 2) the document view displays representations of one or more documents (Fig. 5) and is used when documents are being read and reviewed. The surface is large enough to simulate a user desk and display documents in almost A4 format, like real paper, and can be used individually or collaboratively

In the *wall view* the collection of documents is displayed on the touch screen as tiles on a grid, these tiles can be manipulated using the CX functionality. To produce a natural and intuitive

interaction, we have created specific UI widgets, activated by touch actions, for the document clustering, classification and retrieval functions. The first of these – the ClusteriX button – enables users to choose the number of clusters, e.g. 5, then when activated it groups the documents into clusters and labels the clusters (Stage A), users can then choose to work with a cluster (Fig. 2) and either move directly into document view or further filter the cluster. Filtering is done through a second widget - the “virtual magnet” – which is associated to one of a predefined filtering rule (e.g. filter email documents). The user moves the virtual magnet close to the document tiles and the icons of the documents responding to the query are automatically highlighted and /or moved close to the magnet button (Fig. 3).

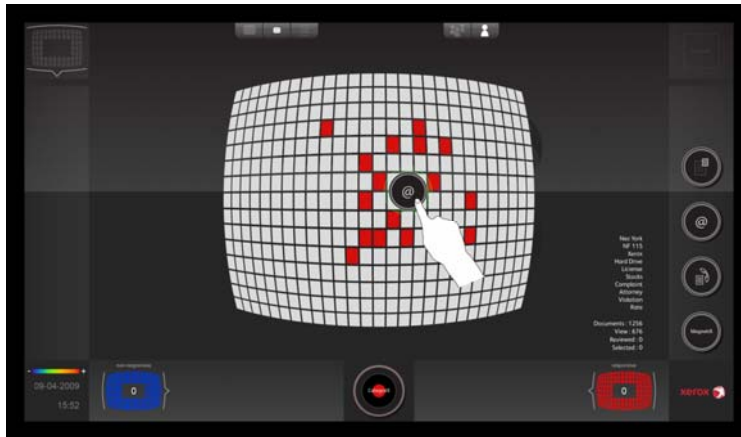


Figure 3: The virtual magnet widget filtering emails from a cluster

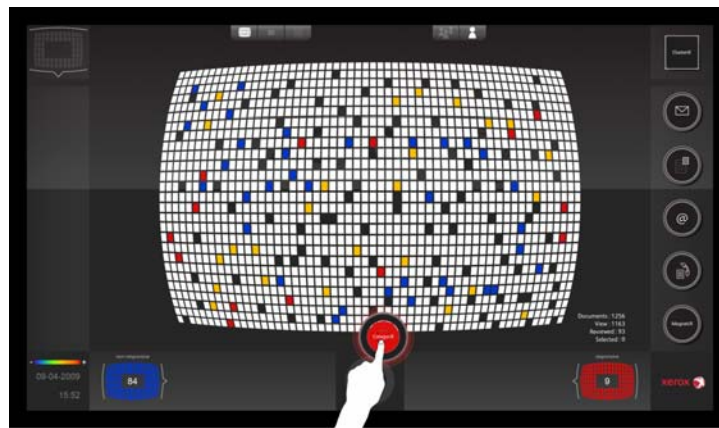


Figure 4: Using the virtual magnet to determine responsiveness

More than one magnet can be used at a time to find subsets of documents (e.g. emails ordered by size parameters). These tasks basically amount to clustering and filtering a particular sub-set of documents out of a larger set, although each of them entails one or more complex algorithms. The complexity of these algorithms is made transparent to the user thanks to the UI widgets. In the same way, binary classification is performed once a model has been created (Stage A or C); in this case, the function associated with the magnet further returns a level of eligibility for each document to be responsive, and the elected documents are displayed around the magnet to different distance reflecting their degree of responsiveness or with different colours: for instance documents most likely responsive are placed closest to the magnet button and are highlighted in red (Fig. 4).

In the *document view* the system presents documents to be reviewed from the subset the user has selected (Fig. 5). Documents can be manipulated through natural gestures, such as rotations,

scaling, moving, etc. As shown in Fig. 5 text can be highlighted, corners turned to indicate hot document status and simple actions move the documents to the responsive/non-responsive bins. The reviewers can also organise their desks almost physically, piling documents, putting documents to one side and moving menu artefacts and buttons to a more convenient place.

The system enables collaborative reviews – the discussion and comparison of difficult documents to improve review accuracy – as several reviewers can stand around the table and manipulate the same set of documents, through duplication and shared highlighting and so on. The touch plate can also become the place where a review administrator can monitor review results and improve performance by inviting reviewers to discuss their tagging decisions, or focusing on hot documents with senior attorneys.

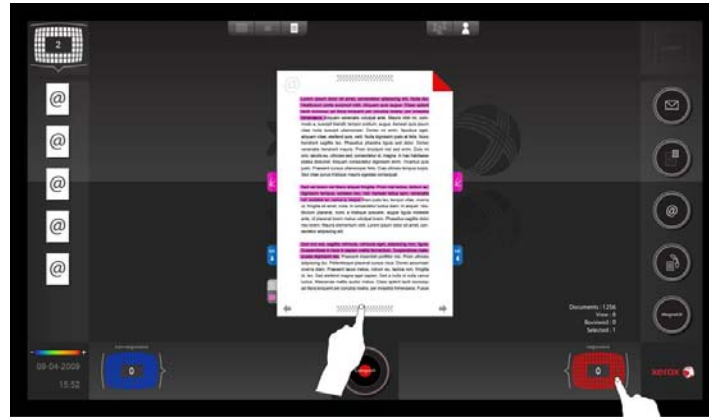


Figure 5: User tagging a responsive document in the document view

4. Discussion

In this paper we have described a prototype system, which uses a novel multitouch interface, built over categorising and clustering technology with the aim of improving the speed and accuracy of document review. The central idea behind the system is to capitalize on the intelligence of reviewers and their progressive understanding of the case: to get them more involved and active in organizing their work and provide them with online feedback through statistical analysis of their work, during and after the review. We believe that, enabling reviewers to interact with sophisticated classification technology through an easy-to-use, intuitive interface which combines the best features of paper and electronic documents will improve the review process. The interface gives the reviewers more autonomy in organising their work as they would like (which pile of documents to review first, do I want similar or different documents at a time, if I reviewed two similar documents differently the opportunity to reflect on that decision). Feedback on performance can help the reviewers learn about the case. The system also enables opportunities for collaborative reviewing to improve accuracy: for example, at certain points during the day, reviewers could be paired to discuss incompatible review decisions; it would improve cross-reviewer accuracy and increase their learning about this case, thus enabling increasingly knowledgeable review. We hope that, as well as providing intra- and inter-reviewer consistency checks, by enabling reviewers to take a more active role in the organisation of their work, DISCO will reduce the monotony of the work and thus its error prone nature. The organisation of documents into groups can increase the speed of review, allaying fears that more autonomy and collaboration might increase review time. Thus the aim is to speed up the review and to improve quality as well as making the work more interesting, thoughtful and intelligent. Indeed, the set up of the system might negate the need to read every document as review counsel gains trust in its classifications of responsive and non-responsive documents. Sampling could be used to test the

documents classified as very likely to be non-responsive and after a period of time providing this was seen to be accurate, only a reduced document set might need to be reviewed – further speeding up the review process.

In terms of the interaction with the multitouch interface, with the design we had an eye to the repetitive nature of the task, so a key design feature was to minimize frequent actions. We employ natural or “almost paper-like” interaction with documents since although litigation support software tools primarily work with electronic documents, paper remains a mainstay of the process in the litigation environment. Even in eDiscovery where large amounts of documents in original digital format are available, paper inputs (e.g. scanning) and outputs (e.g. paper reports) remain frequent. Although behaviours and methodologies are changing, many people, including legal workers, naturally tend to favour paper printouts over digital originals. We believe that the multitouch surface provides the most natural and best interface for DISCO and although it may seem rather expensive at the moment, this technology is bound to drop drastically in price and become affordable for teams of reviewers. We need however to further develop the system, which is currently an un-tested prototype – indeed our rationale for presenting at DESI is to get initial feedback and input on the design. We have already had promising results for the categorizer on legal document sets [12] but at the moment it needs to be used by experts. We hope that DISCO will provide the way forward for use by everyday reviewers, and if it can make their work more interesting at the same time as improving accuracy and speed, that can only be a good thing.

References

1. Greenwood, A. (2007) Attorney at Blah. Washington City paper. <http://www.washingtoncitypaper.com/display.php?id=34054>
2. Blair, D. C. & M. E. Maron (1985). Evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289--299.
3. Hofmann, T. (1999) Probabilistic Latent Semantic Analysis. *Proc 15th Conf. on Uncertainty in Artificial Intelligence*. 289-296. Morgan Kaufmann
4. Gaussier, E., C. Goutte, K. Popat, F. Chen. (2002) A hierarchical model for clustering and categorising documents. *Proc. ECIR-02*. 229-247. Springer
5. Sebastiani, F. (1999) Machine learning in automated text categorization. Tech. Rep. IEI-B4-31-1999 Consiglio Nazionale delle Ricerche, Pisa, Italy
6. Love, N. (2006) Automating Document Review. *CS224n Final Project* http://nlp.stanford.edu/courses/cs224n/2006/fp/natlove-1-natlove_cs224n_final.pdf
7. *The Sedona Conference Journal* (2007) Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery. Vol. 8. 189-223
8. Belkin, N.J. (2008) Some(what) grand challenges for information retrieval. *ACM SIGIR Forum archive*. 42 (1) 47-54.
9. Victor Stanley Inc. v. Creative Pipe Inc. *Civil Action No. KJG-06-2662, 2008 WL 2221841* (D.Md. May 29, 2008). <http://ralphlosey.files.wordpress.com/2008/06/victorstanleymomay29-08final.pdf>
10. Taksa, I, A. Spink & R. Goldberg. (2008) A Task-oriented Approach to Search Engine Usability Studies. *Journal of Software*. 3 (1) 63-73.
11. Multi-Touch G² Touch Screen . PQ Labs, California. <http://multi-touch-screen.net/>
12. Barnett, T., Godjevac, S., Privault, C. Renders, J.M. Wickstrom, R. (submitted *DESI III*) Machine Learning Classification for Document Review
13. Kershaw, A. (2005). Automated Document Review Proves Its Reliability. *Digital Discovery & e-Evidence*. 5. (11)
14. BackStop Software. <http://backstopllp.com/software.html>