Distributed Data Mining for User Sensemaking in Online Collaborative Spaces

Ahmad Ammari
University of Leeds
Leeds LS2 9JT, UK
a.ammari@leeds.ac.uk
+44 113 3431116

Lydia Lau
University of Leeds
Leeds LS2 9JT, UK
l.m.s.lau@leeds.ac.uk
+44 113 3435454

Vania Dimitrova
University of Leeds
Leeds LS2 9JT, UK
v.g.dimitrova@leeds.ac.uk
+44 113 3431674

ABSTRACT

With the increasing growth of data and knowledge intensive online collaborative spaces, such as blogs, wikis, and discussion forums, making sense of what content exists or which posts are the relevant ones for a user to participate in is becoming more difficult, tedious and time consuming. In order to help users make sense of such large-scale information sources, it will be necessary to utilize intelligent technologies and exploit what we know about how people summarize the content, relate the topics, and identify the high-level categories from the huge and diverse information. In this paper, we address this challenge by proposing a novel distributed data mining approach to support user sensemaking. The approach combines the powerful distributed processing of Hadoop Map/Reduce for speed with the scalable data mining of Mahout for dealing with huge volume of data. The output can then be presented to a user in a user-friendly format, such as topic clouds. A pilot case study of an online technical discussion forum was conducted to test the approach.

Author Keywords

Clustering, Collaborative Spaces, Distributed Data Mining, Hadoop Map/Reduce, Mahout, Sensemaking, Topic Clouds

ACM Classification Keywords

H.5.m. Information interfaces and presentation

General Terms

Algorithms, Experimentation, Human Factors

INTRODUCTION

The past few years have observed a phenomenal growth in the Social Web. It has proved to be a popular space for collaboration activities such as upload information, seek advice, provide answers, share experiences, self-expression and networking. The presence of many low cost, easy-to-install Web 2.0 collaborative tools such as blogs, wikis, discussion boards, and instant messaging have tremendously increased the data intensiveness of collaboration on the Web [10].

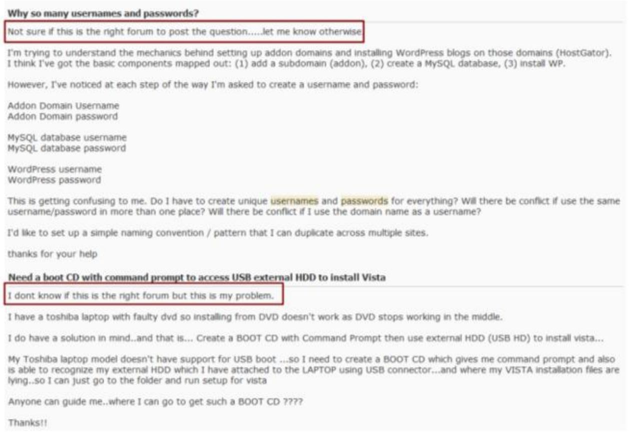
Information overload has become a major problem. While incoming data is rapidly increasing, making sense and filtering what is important for the current situation, becomes difficult and time consuming. This becomes an even bigger problem where collaboration and decision making are taking place. In many cases, the raw information is so overwhelming that users are often at a loss to know even where to begin to make sense of it. “Big Data”, as this problem is called [6], can create stress and cognitive overload to its users [14].

This paper aims to combine techniques from data mining and distributed systems to provide effective support for users to quickly make sense of the huge and dynamic pool of information emerging from the Social Web. Models for individual and collaborative sensemaking are explored to provide a deeper understanding of the potential interaction between users and the supporting technologies.

To test the proposed distributed data mining approach, a case study on discussion forums is being used. To illustrate one of the sensemaking problems in this type of collaborative spaces of the Social Web, let us consider a typical open forum with a large number of categories and discussion threads. Figure 1 depicts two posts written by two different users in the “Computer Assistance” and the “IT Discussion” forum categories in a large online technical support forum, respectively. As highlighted by the red rectangles, both users expressed at the beginning of their posts that they were not sure whether they were posting their technical problems in the right forum category.

There are two main disadvantages for this confusion. Firstly, if a user posts the query in an irrelevant category, those users who have sufficient knowledge to answer the post may not locate it, since they may not browse the posts in that category that is irrelevant to their interests. Secondly, if a user is searching for existing answers for the query, locating them amongst wrongly categorized postings would be challenging. It is proposed that automatic topic identification based on distributed data mining on the content of the discussions will better support the users to make sense of these online collaborative spaces quickly to improve their collaboration.

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.
Copyright 2012 ACM 978-1-4503-1086-4/12/02...$10.00.



Why so many usernames and passwords?

Not sure if this is the right forum to post the question.....let me know otherwise

I'm trying to understand the mechanics behind setting up addon domains and installing WordPress blogs on those domains (HostGator).
I think I've got the basic components mapped out: (1) add a subdomain (addon), (2) create a MySQL database, (3) install WP.

However, I've noticed at each step of the way I'm asked to create a username and password:

Addon Domain Username
Addon Domain password

MySQL database username
MySQL database password

WordPress username
WordPress password

This is getting confusing to me. Do I have to create unique usernames and passwords for everything? Will there be conflict if use the same username/password in more than one place? Will there be conflict if I use the domain name as a username?

I'd like to set up a simple naming convention / pattern that I can duplicate across multiple sites.

thanks for your help

Need a boot CD with command prompt to access USB external HDD to install Vista

I dont know if this is the right forum but this is my problem.

I have a toshiba laptop with faulty dvd so installing from DVD doesn't work as DVD stops working in the middle.

I do have a solution in mind..and that is... Create a BOOT CD with Command Prompt then use external HDD (USB HD) to install vista...

My Toshiba laptop model doesn't have support for USB boot ...so I need to create a BOOT CD which gives me command prompt and also is able to recognize my external HDD which I have attached to the LAPTOP using USB connector...and where my VISTA installation files are lying..so I can just go to the folder and run setup for vista

Anyone can guide me..where I can go to get such a BOOT CD ????

Thanks!!

Figure 1. Two Users Posting Technical Problems in Two Different Forum Categories

The rest of this paper is organized as follows. The next Section introduces user sensemaking and its models, pointing at particular sensemaking operations addressed in this work to support the users in online collaborative spaces. Section 3 presents a novel approach that integrates the Hadoop Map/Reduce distributed data processing framework and Mahout scalable data mining library. In Section 4, the approach is evaluated in a pilot case study using sample content collected from an online technical support forum, producing interesting output whose benefits to the user are illustrated in usage scenarios. In Section 5, we position our work in sensemaking support for collaboration. The paper concludes with a discussion on implications and future work.

USER SENSEMAKING IN COLLABORATIVE SPACES

Sensemaking is an iterative cognitive process that the human performs in order to build up a representation of an information space that is useful to achieve his/her goal [22]. Sensemaking has been used in various fields such as organizational science [25], education and learning sciences [24], communications [8], human-computer interaction (HCI) [22], and information systems [23]. In communications, HCI and information science, sensemaking is broadly concerned with how a person understands and reacts to a particular situation in a given context. Cognitive models that describe the human sensemaking process can be helpful to point at what operations users in collaborative spaces may perform and what support they may need. One particular notional model developed by Pirolli and Card [5], which describes the sensemaking loop for intelligence analysis, helps us to identify particular sensemaking operations that a distributed data mining approach can support in a collaborative environment. The model distinguishes between two cognitive loops of sensemaking operations:

- The foraging loop, which involves operations such as seeking, searching, filtering, reading, and extracting information, and
- The sensemaking loop, which involves operations such as searching for evidence, searching for support, and re-evaluation, which aim to develop a mental model from the schema that best fits the evidence.

The operations involved in the defined loops highlight the importance of two high-level cognitive processes that a user of a collaborative space (e.g. discussion forum) performs: categorization and schema induction [15]. In the foraging loop, the user tries to identify coherent categories, or topics, which summarize the underlying content and aid the user's filtering and searching to find the content relevant to the needs. In the sensemaking loop, on the other hand, the user tries to induce potential high-level schemas, or themes, from the identified topics. This is done by inducing the relations between the topics and evaluating the accuracy of those schemas. For example, if the user **relates** a collection of **identified** topics that include the terms { facebook, twitter, tweets, blogs, wordpress, wiki } to each other, she may be able to **induce** a high-level theme, which is { social media }, since the combination of the preceding topics is highly relevant to that theme.

To help the user quickly make sense of the increasing volume of content in data intensive collaborative spaces, we propose an integrated approach that combines distributed data processing (by using Hadoop Map/Reduce) and large-scale data mining (by using the Mahout machine learning library). The novelty of the approach lies in the exploitation of distributed, scalable data mining processes, particularly data preprocessing and cluster analysis, in order to support the user sensemaking of data intensive collaborative spaces, so the user can quickly:

- **identify** the fine-grained topics that represent the existing content (e.g facebook, twitter, blogs),
- **relate** between the identified topics, to ..
- **induce** the high-level theme these topics represent (e.g social media)

THE PROPOSED DISTRIBUTED DATA MINING APPROACH

In this Section, we introduce a novel approach that integrates the Hadoop Map/Reduce distributed computing model with scalable data mining techniques in Mahout, particularly text cluster analysis algorithms, in order to support the user sensemaking in collaborative spaces, which were highlighted in the previous Section. Figure 2 depicts an overview of the approach, which includes three main phases: Content Pre-Processing, Content Clustering, and Topic Modelling.

When the users collaborate, they create content. This could be in the form of discussion threads, wiki pages, exchanged emails, or instant messages. This content is input to the content pre-processing phase of the sensemaking-support approach, in which it is being prepared for the subsequent phase, Content Clustering, where the pre-processed content is clustered into distinct groups based on the content

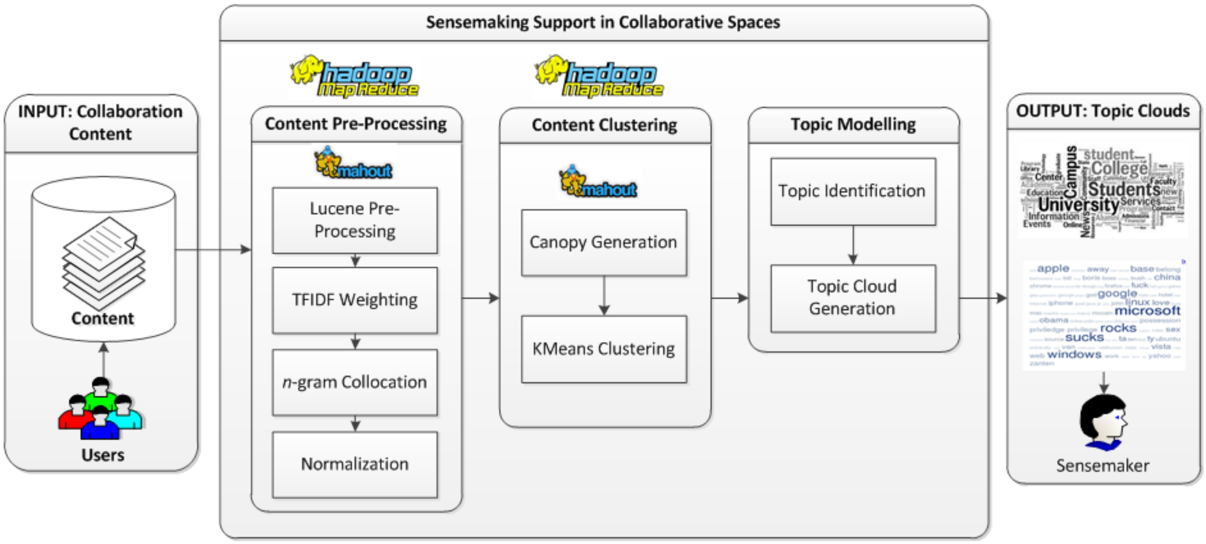


Figure 2. An Integrated Hadoop Map/Reduce and Mahout Distributed Data Mining Approach

similarity. The derived clusters are then provided to the topic modelling phase to extract the key topics and their relative weights from each cluster and generate a topic cloud representation of each cluster, allowing the sensemaking user, or the *sensemaker*, to identify the key topics that the content is all about, relate between them in each topic cloud, and induce the high-level themes these topics represent, thus making sense of the collaborative space.

Adopting Hadoop Map / Reduce and Mahout for Distributed Data Mining

Data intensiveness in online collaborative requires that machines store and process continuously increasing volumes of user-created content. The exponential growth of data first presented challenges to cutting-edge businesses such as Google, Yahoo, Amazon, and Microsoft. Such search and e-Commerce tools needed to go through terabytes and petabytes of data to figure out which websites were popular, what books were in demand, and what kinds of ads appealed to people. Existing tools were becoming inadequate to process such large data sets [16]. Google was the first to publicize Map/Reduce—a system they had used to scale their data processing needs. This large-scale distributed data processing system aroused a lot of interest because many other businesses were facing similar scaling challenges, and it wasn't feasible for everyone to reinvent their own proprietary tool. Apache software foundation[1] saw an opportunity and led the charge to develop an open source version of Map/Reduce called Hadoop[2]. Today, Hadoop is a core part of the computing infrastructure for many data intensive collaborative spaces, such as Yahoo , Facebook , LinkedIn , and Twitter.

The success of Hadoop Map/Reduce motivated many open source communities to develop libraries that can exploit Map/Reduce to run computationally-expensive algorithms. One of the new initiatives is Apache Mahout[3]. Mahout is a scalable machine learning library that implements many data mining algorithms used to solve many data intensive tasks, such as recommendations, clustering, and classification. Although it is still in the development phase, Mahout aims to be the machine learning tool of choice when the collection of data to be processed is very large and dramatically increasing.

The two computationally-expensive phases in our approach, Content Pre-Processing and Content Clustering, are designed to run as Hadoop Map/Reduce jobs, utilizing the Mahout data mining programs in a distributed model.

Content Pre-Processing

Machine learning algorithms, such as cluster analysis, requires that the unstructured, user-created content (e.g. discussions) in the collaborative space be pre-processed into a structured format before being mined. The Content Pre-Processing phase consists of four Mahout components that perform pre-processing of content prior to data mining.

Lucene Pre-Processing

Apache Lucene[4] is a high-performance, full-featured text search engine library written entirely in Java. The Mahout project includes all the text pre-processing class library of Lucene, which has a number of text tokenization, filtration,

[1] http://projects.apache.org/

[2] http://hadoop.apache.org/common/

[3] http://mahout.apache.org/

[4] http://lucene.apache.org/java/docs/index.html

and analysis Java classes that can be used to perform text pre-processing. This component exploits Lucene tokenization and filtration classes to pre-process the content before running the Map/Reduce cluster analysis jobs on the preprocessed content. Lucene classes leveraged in the approach are: (i) **`StandardTokenizer`**: Removes punctuation and splits words at punctuation. Recognizes Internet host names and email addresses. (ii) **`StandardFilter`**: Normalizes terms by removing plural s, S, and periods. (iii) **`LowerCaseFilter`**: Normalizes the term text to lowercase, and (iv) **`StopFilter`**: Removes noisy stopwords that do not contribute to the semantics of the content. Examples are determiners, conjunctions, and prepositions.

n-gram Collocations

Classic TFIDF weighting assumes that terms occur independently of other terms, but vectors created using this method usually lack the ability to identify key features of documents, which may be dependent. To circumvent this problem, Mahout implements techniques to identify groups of terms that have an unusually high probability of occurring together, such as social media, operating systems, and data mining. Mahout allows the creation of document vectors that include both unigrams (single terms) as well as n-gram collocations (multi-term phrases), where n is the maximum number of terms in the phrase. Moreover, Mahout solves the problem of having meaningless bigrams (2-gram phrases) such as “it was”, “was the”, … etc by passing the n-grams through a log-likelihood test, which can determine whether two words occurred together by chance or because they form a significant unit. It then selects the most significant ones and prunes away the least significant ones. The approach uses the **`DictionaryVectorizer`** class.

Normalization

In Mahout’s language, normalization is the process of decreasing the magnitude of large document vectors and increasing the magnitude of smaller document vectors in the term document matrix [19]. In Mahout, normalization uses what is known in statistics as a *p*-norm [9]. For an *n* dimensional document vector *x*, the *p*-norm is given by the equation:

$$\parallel x \parallel_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} .$$

The chosen norm power *p* depends on the type of operations done on the vector. If the distance measure used is the Manhattan distance measure, the 1-norm will often yield better results with the data. Similarly, if the cosine or the Euclidean distance measure is being used to calculate similarity, the 2-norm version of the vectors yields better results. For best results, the normalization ought to relate to the notion of distance used in the similarity metric [19]. With text content in collaborative spaces, the cosine and Euclidean distance measures yield best clustering results [2]. Therefore, the approach normalizes the document vectors using the 2-norm implementation in Mahout.

Content Clustering

Content clustering, or text clustering, is a major technique in text data mining. The text clustering step involves understanding the similarity and dissimilarity between the given text documents, here the user-generated content, and thus dividing them into meaningful groups sharing common characteristics. Good clusters are those in which the members inside the cluster have quite a deal of similar characteristics. After the collaboration content is pre-processed, it can be given to a Mahout clustering implementation. The approach uses two clustering implementation in Mahout to perform content clustering; Canopy Generation and KMeans Clustering.

Canopy Generation

Clustering algorithms implemented in Mahout require that the number of clusters is known before generating the clusters. However, in data intensive problems, the number of clusters, or k, is usually unknown. A number of techniques known as approximate clustering algorithms can estimate the number of clusters and the approximate location of the centroids from a given data set. One algorithm implemented in Mahout is canopy generation [18], which is unsupervised pre-clustering algorithm often used as pre-processing step for the KMeans algorithm. In canopy generation, the input set of points is divided into overlapping clusters known as canopies. Canopy generation tries to estimate the approximate cluster centroids (or canopy centroids) using a fast distance measure and two distance thresholds, T1 and T2, where T1 > T2. when clustering a large collection of text content using the Euclidean distance measure, the values of T1 and T2 are usually large (e.g 2000, 1500 respectively). In Mahout, the canopy generation algorithm is executed as a Map/Reduce job using the **`CanopyDriver`** class. The number of generated canopies can then be fed into the KMeans clustering algorithm as the approximate number of clusters k.

KMeans Clustering

The most stable clustering algorithm in Mahout is the well-known KMeans algorithm, which uses the Euclidean distance measure to cluster the term document matrix using the Euclidean distance into k distinct clusters, where k is the input number of clusters [1]. The feature weighting KMeans algorithm is used to generate the clusters of user-generated content based on text similarity. KMeans has been used successfully to solve the problem of clustering large and complex text data with good results [13]. The approach uses the Mahout KMeans version implemented to run as a Map/Reduce job using the **`KMeansDriver`** class, clustering the collaboration content into k clusters, where k is pre-determined by the previous canopy generation component.

Topic Modelling

In the Topic Modelling phase, the approach employs the content clusters generated by KMeans to identify the most frequent terms that represent the topics of the content that belongs to each cluster and visualize the terms to the sensemaker. In the Topic Identification component, the approach identifies the topics of each cluster by querying the top n terms having the maximum weights in the cluster centroid, which is a vector of the average TFIDF weights for all the terms and bigrams used in the clustering process. The Mahout launcher program **`clusterdump`** is used to query the top terms in the cluster centroids. The retrieved terms from each cluster centroid and their weights are then fed into a Tag Cloud creator component to generate a Topic Cloud representation for each cluster. By depicting the generating clouds, the sensemaker can identify the topics each cluster of content consists of, relate between the topics, and induce the main theme these topics belong to. The Java-based OpenCloud[5] tag cloud class library is used by the Topic Cloud Generation component to generate the topic clouds.

TECHNICAL DISCUSSION FORUMS: A CASE STUDY

To evaluate how the approach can help users make sense of the collaboration content, a pilot case study has been conducted on online technical discussion forums. In the following sub-sections, we describe the forums from which we collected the data, evaluate the need of the approach, determine which content to mine by the approach to derive the topic clouds, and present example experimental output.

A Public Technical Forum

WebProWorld[6] is a large online discussion forum for IT support. It has a number of predefined subforums, each containing a number of discussions that focus on a particular discussion category. Currently the subforums are “computer assistance”, “search engines”, “webmaster, IT and security”, “e-Commerce” and so on. Each discussion in a subforum is a series of postings linked by a thread. Members use the forum to seek help from each another and exchange ideas, tips, news, and information. However, as discussed in the first Section, users may find it hard to identify the right subforum (category) to post new discussion threads, involve in on-going discussions of interest, or search for answers to specific problems. This may be due to one or more of the following reasons:

1. The defined categories in the forum represent the perspectives of the forum designer(s) on the appropriate grouping of discussion topics. This top-down approach may not take into account the perspectives of the collaborating users and their contributions.
2. As technology evolves rapidly, new topics may become trending categories for discussion. These germinating categories could be embedded across established categories. As a result, a user may not be able to find the right thread to contribute to one of these new hot topics if the top-level categories are not kept up-to-date.
3. Some predefined categories may be very broad. For example, the “IT Discussion” category is a very broad label for a sub-forum, containing over 3000 threads and over 8000 individual posts. While it could be a ‘catch-all’ category, a user may need to spend considerable time to scan through the threads in order to decide where to post his/her specific IT problem in this broad sub-forum, or to guess which other specific sub-forums would be more relevant to the problem.

The above illustrates the potential cognitive overloading in making sense of the discussion forum. During the foraging loop, the sensemaker may miss relevant discussion category or may need to spend a long time to read through the discussions in order to extract the main trends. During the sensemaking loop, the sensemaker may spend significant time and effort to derive relationships between discussions and their individual topics so that the main discussion themes could be induced.

This case study serves as an experiment to test how the integrated approach can better support the aforementioned sensemaking operations.

Objectives for the experiment

It is proposed that better sensemaking support can be provided by the deployment of distributed data mining approach. There are three main objectives for the experiment:

1. Investigate the extent of sensemaking support needed for the public technical forum.
2. Determine which content representation for clustering is more appropriate to derive topic clouds for the sensemaker.
3. Illustrate how the output of the approach could provide sensemaking support.

Sample Input

Four subforums were chosen for the experiment:

- Two subforums representing fairly specialized categories – **SEO** (Search Engine Optimization) and **e-Commerce**;
- Two subforums representing broad categories – **IT** and **Computer Assistance**.

From each subforum, 50 contributions were randomly sampled to allow equal representation.

Experimental Methods and Result Analysis

Main steps and outcome of the experiment are discussed under each objectives as stated above.

[5] http://opencloud.mcavallo.org/

[6] http://www.webproworld.com/webmaster-forum/forum.php

Examining the Need to Support User Sensemaking

To address the first objective, Mahout's implementation of Content Pre-Processing and the KMeans clustering in Map/Reduce was used to preprocess and cluster the 200 discussions into four distinct clusters (a number chosen to equal the number of subforums used). We argue that if the categories in the forum were sufficient to organize similar discussions together, the derived clusters should reflect this fact by providing a close match between one cluster to one subforum.

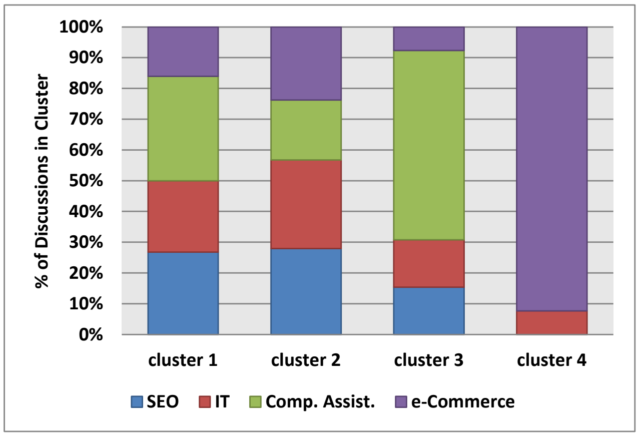


Figure 3(a). Distribution of Four Categories in Four Mahout-based Clusters by Title

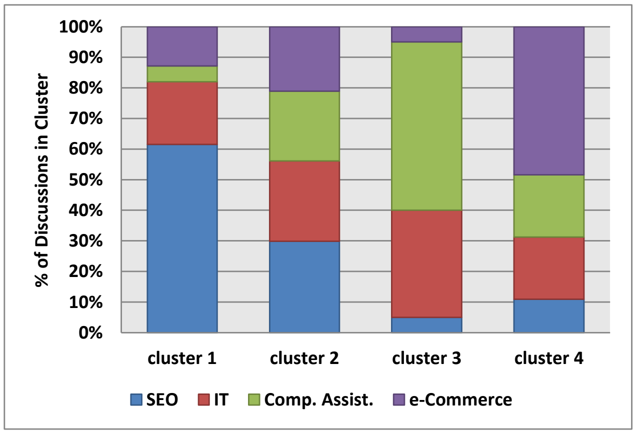


Figure 3(b). Distribution of Four Categories in Four Mahout-based Clusters by Title and First Post

Figure 3 shows the percentages of each of the four pre-defined categories in each derived cluster after clustering the titles of the discussions (3a), and clustering the titles and first posts (3b). The following observations were made:

(i) In Figure 3(a): Out of the four clusters, two clusters did not have any clear one-to-one mapping with the categories. For clusters 3 & 4, each was dominated by a particular category (cluster 3 - more than 60% from the Computer Assistance category and cluster 4 - 90% from the e-Commerce category). This supports our claim that better sensemaking mechanism is needed to enable a better match between catergories and content.

(ii) Clustering by using the titles and content of first posts further confirm our claim. As shown in Figure 3(b): For clusters 1, 3 & 4, each had a marginal dominant category but still had a large proportion of discussions from other three categories. Cluster 2 had an equal mix of discussions from all four categories. This demonstrated a clear need for assisting the users to be quickly aware of the topics being discussed and to locate the right place for posting and reading.

Determining the Right Content Representation

To address the second objective, two cluster validity measures were used to examine which of the two content representations, (i) titles only or (ii) titles and first posts, provided better clusters. The cluster validity measures used are:

1. **Davies-Bouldin Index (DBI):** DBI [7] is a well-known metric for determining how well the content (in this case, discussions) are clustered by the algorithm. For each cluster, the DBI is the ratio between two calculated measures, the measure of discussion scatter within the cluster and the measure of separation between the cluster and the other clusters in the model. Discussions that are less scattered to each other in one cluster and more separate between different clusters have more common terms between them, thus focus on similar topics. Therefore, the smaller the first measure and the bigger the second measure, the better the clustering of discussions is. The DBI for the whole model is the average value of the DBI values for all the clusters in the model. The smaller the average DBI, the better the model is for achieving a coherent set of similar discussions.
2. **Item Distribution Measure:** Item distribution is a metric that takes the number of discussions in each cluster (or cluster density) into account when examining the derived clusters based on the methods of density-based clustering [12].

   Given a total number of clusters N that contains a total of n discussions, the item distribution measure is given by the equation:

$$\sum_{i=1}^{N}\left(\frac{c_i}{n}\right)^2$$

   where c_i is the number of discussions that belong to cluster i. For a situation where one cluster dominates and the other clusters are smaller in comparison, this value will be closer to 1.0. For a situation where the clusters have a relatively equal number of discussions, the value tends to be 1/N, hence closer to 0.0. A clustering model that has an item distribution value closer to 1.0 will derive minor distinct clusters with topic-specific discussions. This helps the user to locate a topic-specific discussion easier as the minor clusters having such discusisons will be made explicit in topic clouds.

For each representation, ten KMeans clustering models (from k = 3 to k = 12 clusters) were developed and run as Map/Reduce jobs on a single-node Hadoop cluster using the Mahout KMeans implementation. Because Mahout currently does not implement cluster validity measures, RapidMiner[7] was used to calculate the average values of the DBI and item distribution measures for all the clustering models. The following observations were made:

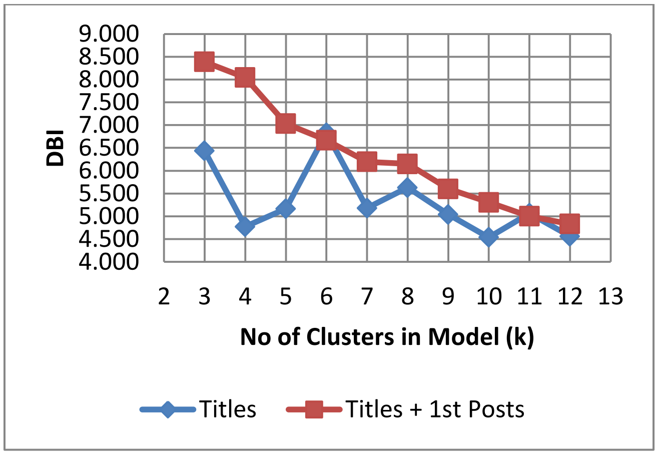


Figure 4(a). Davies-Bouldin Index Values for Two Content Representations Used

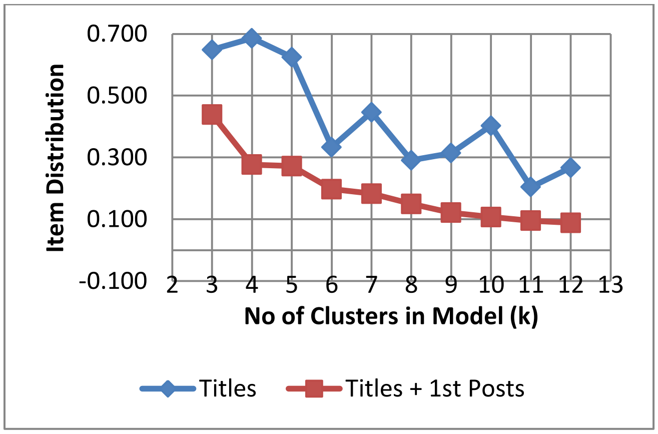


Figure 4(b). Item Distribution Values for Two Content Representations Used

(i) From Figure 4(a), one can conclude that the representation using titles only is better than the representation which used both titles and first posts. The lower DBI values mean that the discussions within the derived clusters are higher in similarity and well separated from a cluster to another. Hence visualizing the topics as a cloud per cluster using title-only representation will be more effective in helping users to spot discussion themes.

(ii) Figure 4(b) confirmed the preference to using title-only representation. It shows that all ten models which clustered discussions with title-only representation have item distribution values closer to 1.0 than the models which clustered discussions with title and first post representation. As a result, clustering discussions represented by titles only has more ability to derive minor distinct clusters that help a sensemaker to discover discussions that might have been buried in broad subforums.

Exploitation of the output for User Sensemaking

To address the third objective, the proposed distributed data mining approach was applied. Title-only representation was being used and 'k' is set to 10 so that discussions are grouped into 10 clusters. The Topic Modelling component in the approach identified the terms and bigram phrases having the highest cluster centroid weights in each cluster and used them to build a topic cloud per cluster.

Figure 5 depicts the most important terms and bigram phrases in three of the built 10 topic clouds, where each cloud presents the topics of discussions in a derived cluster. The following three usage scenarios illustrate how these topic clouds can better support user sensemaking by comparing with a current practice which has been empirically checked.

<u>Scenario 1 - Use Predefined Categories to find Relevant Content:</u>

- **Context:**

John is a businessman who is used to make backup copies of his important job-related data on his HP local server. John is used to use specific HP software, namely Disaster Recovery, to make incremental backups of his business documents every few days. Unfortunately, John's copy of Disaster Recovery has become corrupted and HP has stopped supplying the software. John is not aware of any good alternative software to Disaster Recovery and wants to get some advice from experts.

- **Current Practice:**

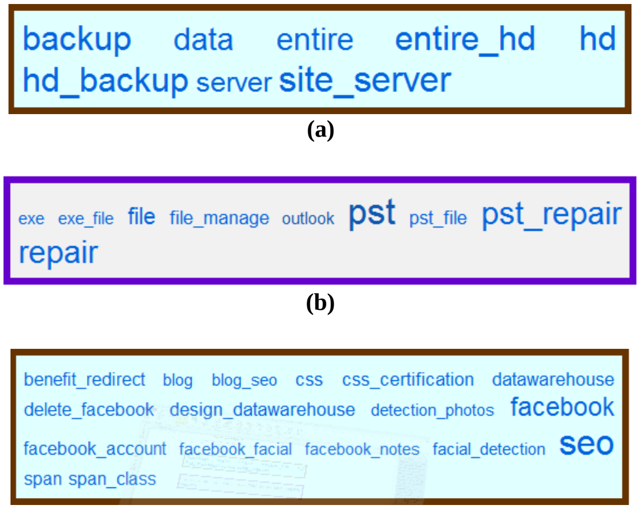
John logged onto a technical discussion forum.. He firstly scanned through all the available categories in the forum, which are "Search Engine Optimization", "IT", "Computer Assistance", and "e-Commerce". John decided that he wants to browse the existing discussions in the "Computer Assistance" category since he was seeking assistance on how to backup data on his HP server's hard disk. However, the "Computer Assistance" category is a very broad category in which he finds over 500 existing discussion threads. John spent around 2 hours browsing the discussions and tried to find a user who posted a similar problem and got useful answers, but he could not find any.

- **Using the Topic Clouds:**

John logs onto the technical discussion forum. Instead of the predefined categories, the forum site displays a number of topic clouds, each cloud depicts the most frequent topics that represent the discussions of specific themes. Instead of

[7] http://rapid-i.com/content/view/181/190/

reading the content of discussions, John now can rapidly visualize the topics inside each cloud and determine the relevance of the discussions represented. A particular cloud ((a) in Figure 5) has drawn John's attention as it includes topics such as "backup", "data", "hd" and "hd_backup". John induces that the discussions belonging to this cloud are mainly about hard disk (hd) backup. John clicks on the bigram phrase "hd_back" in the cloud and retrieves a list of 8 discussions that belong to this cloud. Amongst this much reduced set of discussions, John is able to rapidly locate a particular discussion thread in which a member was asking for good alternatives to the Disaster Recovery backup software. John reads the replies and obtains a recommendation by a couple of fellow members. The whole process takes John a much shorter time. It also happened that the question was posted in the "IT" category, and not in the "Computer Assistance" category as John thought in the current practice.


backup data entire entire_hd hd hd_backup server site_server
exe exe_file file file_manage outlook pst pst_file pst_repair repair
benefit_redirect blog blog_seo css css_certification datawarehouse delete_facebook design_datawarehouse detection_photos facebook facebook_account facebook_facial facebook_notes facial_detection seo span span_class

(a)

(b)

(c)

Figure 5. Sample Topic Clouds Generated from the Derived Clusters by Distributed Data Mining

Scenario 2 - Use Keyword-search to Locate Answer to a Problem

- **Context:**

Mary is a sales representative in a big commercial company and she is a heavy user of Microsoft Outlook for exchanging emails with her clients. Unfortunately, her Outlook software suddenly stopped working properly, displaying an error message that "Outlook pst file corrupted due to exceeding the 2 GB .pst file size limit".

- **Current Practice:**

Mary logged onto a technical discussion forum. Mary firstly looked at the predefined categories in the forum to see if any is relevant, such as "MS Outlook". Mary found two topic-specific categories, "Search Engine Optimization" and "e-Commerce", which are irrelevant to MS Outlook problems, and another two broad categories, "IT" and "Computer Assistance", which contain hundreds of discussion threads. Since none of the predefined categories are helpful, Mary decided to use keyword search. She entered the keyword "pst", which triggered a response from the forum, saying that the words she used in her search are either very common, too long, or too short. Mary tried "Outlook" as the keyword, which returns 116 discussion threads. After selecting five threads to read from the search result based on the thread titles, she finally found a thread that discusses a similar problem to hers. Mary spent around half an hour to find the most relevant discussion.

- **Using the Topic Clouds:**

Mary logs onto the technical discussion forum. Like John in the previous scenario, she can now visualize the main topics of the diverse discussions that exist in the forum using the topic clouds. Mary's attention is drawn rapidly by a particular cloud ((b) in Figure 5) that includes very relevant topics ("pst", "pst_repair", "pst_file", "outlook"). Mary clicks on the bigram phrase "pst_repair" in the cloud. A list of 12 discussions is returned, one of which is the same most relevant thread that she spent half an hour in the current practice to find, but with much less time and effort.

Scenario 3 - Use the Forum to Discover Emerging Technical Areas

- **Context:**

Mark is a course developer in an IT training company. He has been asked to update the training resources on IT support which will reflect current demands. Mark wants to identify the emerging areas in IT and the kind of hot topics that people seek help from support staff.

- **Current Practice:**

Mark logged onto a technical discussion forum. To discover the emerging areas in IT, he couldn't rely on the category titles predefined by the forum designers since these titles are static as Mark used to see them every time he used the forum. To identify emerging topics, Mark used an advanced search feature to retrieve discussions that occurred over the last month. It returned 107 discussions. Mark then manually went through all the discussions in order to summarize the key trends which took him several hours to accomplish.

- **Using the Topic Clouds:**

Similar to the previous scenarios, a list of topic clouds are presented to Mark once he logs onto the discussion forum. Mark can use the topic clouds to direct his effort for further investigations. For example, some of the emerging topics which Mark observes in a particular cloud ((c) in Figure 5) are very fine-grained specific topics that none could be deduced from existing category titles. These are:

1. **Photo features in social networks:** ("facebook", "facebook_facial", "facial_detection", "detection_photos").
2. **Optimizing Search Engines for Blog Search:** ("blog", "blog_seo").

3. **Design of Datawarehousing Systems:** (“datawarehouse”, “design_datawarehouse”).
4. **Certificates and Skills in Web Design:** (“css_certification”, “span_class”).

The topic clouds do not remove the need for Mark to do the association and summarization, but this can now be done in a more focused manner and in a shorter time.

RELATED WORK

There are existing tools that have been designed for individuals to support human sensemaking on either large document collections or the web. SenseMaker [3] supports information exploration tasks by enabling users to search multiple, heterogeneous sources of information. Entity Workspace [4] helps users make sense of large document collections by enabling automatic highlighting of important terms, note-taking with an electronic notebook, importing text from documents, adding comments, and organizing information. The Sensemaking-Supporting Information Gathering [21] system supports sensemaking in web search tasks. The user searches information on the web and organizes the information gathered into a hierarchical tree structure. ScratchPad [11], developed as an extension to the standard browser interface, assists users in making sense of information found on the web. However, processing large content from the Web to support sensemaking can be computationally-expensive and time-consuming, which hinders the sensemaking process. We add to the aforementioned works by addressing the intensive data processing challenge using distributed computing technologies to efficiently support sensemaking of large-scale data.

NLP techniques were used to extract more meaningful and higher quality political opinions from micro-blogs (e.g. user-generated content on Twitter). A very recent work of Maynard and Funk [17] mined positive, negative and neutral sentiments from tweets to make sense of the public opinions regarding specific political subjects. Our approach further extends mining user-generated content to detect the key topics that people tend to talk about during their collaboration.

There have been initial attempts to exploit data mining and machine learning techniques to support human sensemaking. [15] used graph mining techniques to assist users in organizing and understanding large collections of information. In that work, the different documents for an author have been clustered based on their linkage to each other in a graph representation. Our approach extends this idea into two main aspects. Firstly, the digital traces that users create in collaborative spaces can be similar in content without being explicitly linked to each others. We observed this in our experiments which clustered similar discussions from different unconnected categories. Secondly, we adopted Mahout in Map/Reduce implementations of machine learning to address the computationally-expensive processing challenge that statistical learning algorithms, such as clustering, require to analyze the intensive data in collaborative spaces.

CONCLUSIONS AND FUTURE WORK

Contemporary spaces of collaboration in the Web 2.0 era are often associated with huge and continuously increasing amount of user-created content. Sensemaking is becoming a crucial first step for effective collaboration and decision making in this data intensive environment. Our proposed novel approach is based on cutting-edge distributed computing and scalable data mining technologies to support the rapid sensemaking in collaborative spaces which consist of dynamically changing user contributed content. By clustering, topic-specific groups of content, which were previously buried in pre-defined broader categories, can be generated. The additional visualization as topic clouds enables sensemaker to easily and quickly identify the topics of interest, relate between them, and induce the main theme of the topics.

It is acknowledged that Mahout is still at an early development stage. Our experience has highlighted the following areas for further development:

- **Cluster validity measures:** We had to use Mahout to perform Pre-Processing and Clustering, then switched to RapidMiner to compute the DBI and Item Distribution measures. Support for these well-known cluster validity measures in Mahout will provide a seamless progression from one stage to another.
- **Advanced dimensionality reduction algorithms:** text clustering often faces the curse of high dimensional data. This can be addressed using dimensionality reduction techniques such as Latent Semantic Indexing (LSI) [20], which is a pre-clustering technique that detects correlational attributes and merges them into a single feature. Due to the reduction in dataset dimension, clustering becomes faster. It also improves the quality of clustering since the merged features dominate the clustering process. Mahout is highly encouraged to implement dimensionality reduction with LSI.
- **GUI support:** Similar pre-processing and machine learning processes can be applied faster in GUI-supported data mining platforms, such as RapidMiner. Mahout currently lacks this feature. This increased the approach development time and effort. A GUI plug-in of the Mahout class library into well-known Integrated Development Environments is highly desirable.

As future work, we aim to further develop this approach into a sensemaking-support service which could be plugged-in as a service or a widget in any collaborative spaces. Further user trials will be needed to test the transferability of the approach in different domains.

ACKNOWLEDGMENTS

This publication has been produced in the context of the EU Collaborative Project "DICODE - Mastering Data-Intensive

Collaboration and Decision" which is cofunded by the European Commission under the contract FP7-ICT-257184. This publication reflects only the author's views and the Community is not liable for any use that may be made of the information contained therein.

REFERENCES

1. Abonyi, J. Feil, B. Cluster Analysis for Data Mining and System Identification, Birkhäuser Verlag AG, Berlin, (2007)
2. Agarwal, N., Liu, H. Modelling and Data Mining in Blogosphere, In: Synthesis Lectures on Data Mining and Knowledge Discovery, R. Grossman, ed., Morgan & Claypool Publishers, vol. 1, (2009)
3. Baldonado, M.Q.W. and Winograd, T. SenseMaker: An information-exploration -interface supporting the contextual evolution of a user's interests, In: Proc. CHI 1997, (1997), 11-18
4. Billman, D. and Bier, E.A. Medical Sensemaking with Entity Workspace, In: Proc. CHI 2007, (2007), 229-232
5. Card, S. and Pirolli, P. The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis, In: Proc. International Conference on Intelligence Analysis, (2005)
6. Cukier, K. A special report on managing information: Data, data everywhere. In: The Economist, February 25, (2010)
7. Davies, D., and Bouldin, D. W. A Cluster Separation Measure, In IEEE transactions on Pattern Analysis and Machine Intelligence, PAMI-1 Issue:2, (1979), 224 – 227
8. Dervin, B. From the mind's eye of the user: The Sense-Making qualitative-quantitative methodology, In: Sense-Making methodology reader: Selected writings of Brenda Dervin. Hampton Press Inc, Cresskill, NJ, USA, (2003)
9. Elden, L. Matrix Methods in Data Mining and Pattern Recognition, Society for Industrial and Applied Mathematics, SIAM, Philadelphia, PA, USA (2007)
10. Fichter, D. The many forms of e-collaboration: Blogs, wikis, portals, groupware, discussion boards, and instant messaging. In: Online 29, 4 (2005), 48--50
11. Gotz, D. The ScratchPad: sensemaking support for the web, In: Proc. of WWW, (2007), 1329-1330
12. Hans-Peter, K., Peer, K., Jörg, S., Arthur, Z. Density-based clustering. In: WIREs Data Mining Knowledge Discovery, vol. 1, doi: 10.1002/widm.30 (2011), 231-240
13. Jing, L., Ng, M., Xu, J., and Huang, J. Z. Subspace Clustering of Text Documents with Feature Weighting K-Means Algorithm, In: Lecture Notes in Computer Science, Vol. 3518, (2005), 802-812
14. Kirsh, D. A Few Thoughts on Cognitive Overload, In: Intellectica, vol. 1, no. 30, (2000), 19-51
15. Kittur, A., Chau, D.H., Faloutsos, C. and Hong, J. I. Supporting Ad Hoc Sensemaking: Integrating Cognitive, HCI, and Data Mining Approaches, In: Sensemaking Workshop at CHI, Boston, M.A (2009)
16. Lam, C. Hadoop in Action. Manning Publications (2011)
17. Maynard, D., and Funk, A. Automatic detection of political opinions in tweets. In: Proc. of MSM 2011: Making Sense of Microposts Workshop at 8th Extended Semantic Web Conference, Heraklion, Greece, May (2011)
18. McCallum, A., Nigam, K., Ungar, L.H. Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, (2000)
19. Owen, S., Anil, R., Dunning, T., and Friedman, E. Mahout in Action, Manning Publications (2011)
20. Papadimitriou, C. H., Raghavan, P., Tamaki, H., and Vempala, S. Latent semantic indexing: A probabilistic analysis. In: Journal of Computer and System Sciences, 61:2, (2000), 217-235
21. Qu, Y. A. Sensemaking Supporting Information Gathering System, In: Ext. Abstracts CHI 2003, (2003), 906-907
22. Russell, D.M., Stefik, M.J., Pirolli, P., and Card, S.K. The cost structure of sensemaking. In: Proc. SIGCHI, ACM Press New York, NY, USA (1993), 269-276
23. Savolainen, R. The Sense-Making theory: Reviewing the interests of a user-centered approach to information seeking and use, In: Information Processing and Management, 29, 1, (1993), 13-18
24. Schoenfeld, A.H. Learning to think mathematically: Problem solving, metacognition, and sensemaking In mathematics, In: Handbook for Research on Mathematics Teaching and Learning, MacMillan, New York, NY, USA, (1992)
25. Weick, K.E. Sensemaking in Organizations, In: Sage Publications Inc, Thousand Oaks, CA, USA, (1995).