# Novel Applications and Research Problems for Sensor-Clouds

Yu-Hsn Liu[1], Kok-Leong Ong[2], Andrzej Goscinski[2]
[1]Department of Computer Science and Computer Engineering
LaTrobe University, Bundoora, Victoria 3086, Australia
[2]School of Information Technology
Deakin University, Burwood, Victoria 3125, Australia
y65liu@students.latrobe.edu.au, {leong, ang}@deakin.edu.au

***ABSTRACT:*** *Recent developments in sensor networks and cloud computing saw the emergence of a new platform called sensor-clouds. While the proposition of such a platform is to virtualise the management of physical sensor devices, we foresee novel applications being created based on a new class of social sensors. Social sensors are effectively a human-device combination that sends torrents of data as a result of social interactions. The data generated appear in different formats such as photographs, videos, or short texts, etc. Unlike other sensor devices, social sensors operate on the control of individuals via their mobile devices like smart phones, tablets or laptops. Further, they do not generate data at a constant rate or format like other sensors do. Instead, data from social sensors are spurious and varied, often in response to social events, or a news announcement of interests to the public. This collective presence of social data creates opportunities for novel applications never experienced before. This paper discusses three such applications utilising social sensors within a sensor-cloud environment. Consequently, the associated research problems are also presented.*

**Keywords:** Sensor-cloud, Mobile applications, Social networks

**Received:** 17 January 2012, Revised 12 March 2012, Accepted 15 March 2012



## 1. Introduction

In recent years, we see a number of technological advances that converges towards improving '*user experience*'. On the telecommunication front, data is now transmitted through the air with significant speed and capacity. On the storage and computing front, we have faster processors and bigger storage to allow better user interfaces. The combination of a smaller device footprint, better computing power and mobile Internet connectivity has enabled the development of many new technologies such as smart phones, netbooks, tablets, IP cameras, RFID devices, and physical sensors. In turn, these devices created new computing paradigms (and applications) through the way they capture or generate information [1]. This saw the development of sensor networks [2] as a way to collage the different sensor devices into a cooperative network. Likewise, the differences in computing and storage technologies saw the development of cloud computing as a way to simplify the management of complexity and computing resources.

These isolated developments are now converging, in part due to its maturity, but more due to user demands for a seamless-integrated experience across all computing platforms. In turn, the consolidation of a common platform drives the development

of new applications. One such convergence that has a huge potential is the development of a sensor-cloud. Currently, sensors utilise cloud services but the two technologies are at the moment, lacking a powerful common platform. We argue that such a common platform will elevate the user experience to a higher level. To appreciate why this is so, one can look at the computing devices that are now increasingly equipped with sensory elements. Speed cameras, Internet fridges, vending machines, transducers, GPS and proximity detectors are all sensor devices. They are called as such because these computing devices all carry some form of ability to read the surrounding. An empty vending machine for example, is capable of knowing that it is empty and needs a refill. If it is connected to the telecommunication network, it can automatically request a refill. A sensor thus presents opportunities for creating smart applications.

On the other hand, a sensor like a smart phone is capable of generating a lot of data through its sensor elements such as the camera and GPS. A human user can use a smart phone to make videos and upload on YouTube, send a tweet on Tweeter, or take a picture to post on FaceBook. A sensor thus also generates a lot of data. And by a lot, we are talking about huge amount of data generated very quickly within a short time span. To put this in perspective, Bloomberg and Teradata have reported that the amount of data generated by sensors (of all sorts) in the last three years is more than in the past 40,000 years [15]. The combination of smart application opportunities and large volume of data will drive an eventual need for a seamless-integrated cloud platform. Such a sensor-cloud will simplify the access of different devices, and manage the huge volume of data transiting across the different computing networks and storage.

In addition to simplify access, the sensor-cloud will be able to broke a huge volume of computing resources in a manner that is easily accessible. As accessibility is achieved, the seamless-integrated experience becomes possible. This will be realised through the sensor-cloud as a set of context-aware applications [3]. Gartner predicted that such applications will be huge and will be a key driving force shaping the IT industry in the near to middle term. In this paper, we consider the possible novel applications that may arise out of a sensor-cloud platform. From such possible applications, we discuss the potential research issues that needs solutions. As such, we have organised our paper as follows. In Section 2, we present an overall architecture of a typical sensor-cloud platform. This will provide the background knowledge necessary for our discussion in Section 3, where we present the applications, the research problems and a discussion of the role of the sensor-cloud in enabling such applications. We then conclude in Section 4 with a discussion of the work we are currently undertaking.

## 2. Sensor-Cloud Platform as a Service

The concept of a sensor-cloud is to be a platform providing services as well as collating social sensor data. As such, it is a step beyond a typical cloud providing storage and compute resources. Rather, the sensor-cloud also provides data it collected from sensors, as well as data processing services to enable easier consumption of sensor-data. To simplify consumption of sensor-data, we proposed incorporating additional software services such as data processing services mentioned in Example 1 (Section 3.1) as well a standardised analytical services such as those used in Example 2 (Section 3.2).

To provide a background of what the sensor-cloud will conceptually look like, we extend the architecture of CloudMiner [19]. The CloudMiner is a cloud architecture that provides storage and computer services as well as analytical services. As such, it makes sense to extend this platform to incorporate the virtualisation of sensors, the capture of its data and to incorporate the data processing service mentioned. Figure 1 shows the CloudMiner architecture, which is really an ensemble of clouds, where each cloud perform a specific function.

The Data-Cloud for example, is primarily responsible for data management services and the MiningCloud exposes analytical tools as analytical services. In Figure 1, the changes made to the original CloudMiner are the addition of the SensorCloud to virtualise the sensors as logical devices; the inclusion of data stream services in the DataCloud; and deploying the event-driven model in the AcessPoint and the data stream analytics (DSA) in the MiningCloud. These extensions and its relationship are also shown in the Figure. The discussions that follow will be set on such a platform, where the applications can run in such an ensemble of clouds, tapping into the services it provides.

## 3. Applications and Problems

We present here three novel application ideas that will benefit from a sensor-cloud such as the one depicted in Section 2. Our discussion focuses on them drawing social sensor data from such a platform and thus, allowing us to assume that the complexity of acquiring such data in an adequate format has been already dealt with. These applications thus present themselves as seed

ideas to what could be possible and of potential interests to the research community for futher investigation. We also believe that it will be of interest to the industry looking to extend their current application with an innovative edge.

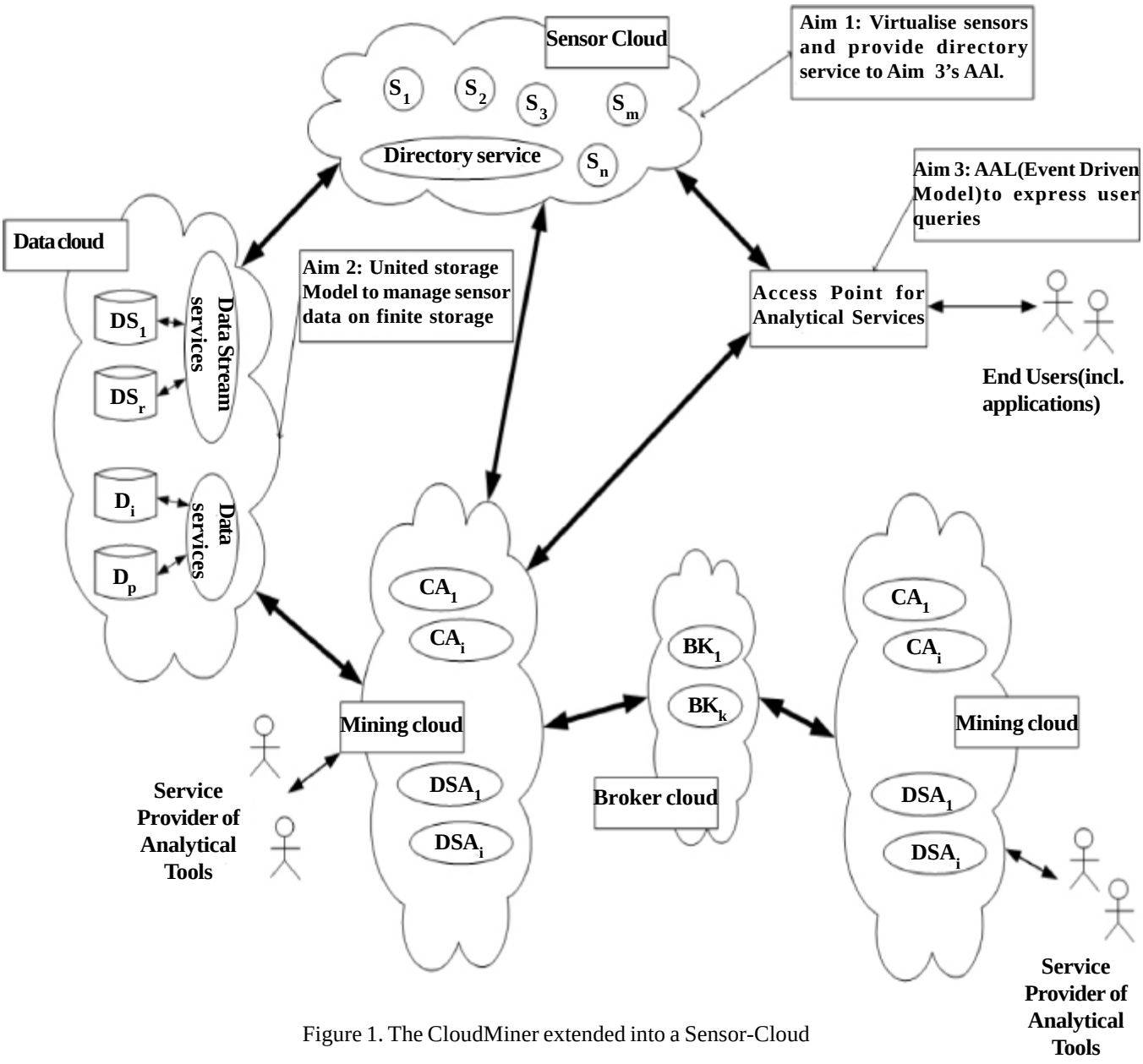


Figure 1. The CloudMiner extended into a Sensor-Cloud

### 3.1 Context Aware Target Marketing

The first application we present has the potential for a novel solution to target marketting on sites like YouTube. As a popular video sharing site centred around user-generated content, users are primarily the drivers of video uploads, comments and ratings. The interaction as a result carries social networking elements that differentiates it from other video sharing sites. Interesting though, the success of monetising YouTube [4] hasn't been as significant as Gmail despite the same owner, i.e., Google. While Google has successfully analysed emails to bring up targeted ads, the AdSense technology wasn't obvious on YouTube. So while YouTube is generating more than 100 million views and producing more than 65,000 video uploads a day [5], the largest biggest broadcaster by eyeballs and content is reportedly missing out on advertising dollars. Clearly, unlike emails where texts can be easily analysed, it is a lot more challenging to identify specific objects in a video. An email for example can mention a specific brand in the text that could be easily picked up. The same specific mention of the brand in a video however

cannot be easily identified without human intervention. The consequence is that users watching generic videos such as 'CSI' are shown ads that appear untargeted.

From the advertising perspective, the YouTube experience is similar to traditional television programmes. Clearly, we are not exploiting the characteristics of YouTube to maximise its effectiveness. As faster Internet speed enables video on demand, more users will move away from traditional television. As a matter of fact, earning reports in Australia and the anecdotal evidence of a drop in viewerships suggest that this is already happening [6, 7]. As time poor young adults with both earning and spending power turn towards sites like YouTube, targeted marketing technology will be the key to enable significant selling opportunities on YouTube. The sensor-cloud model presents such opportunities in the form of accessing analytical software as a service. As users post comments in response to their video views from their mobile sensor-enabled devices, analytical software could be used to identify relevant keywords for showing the advertisement with the most potential for a click-through. The innovation here is the strategic combination of location information from the mobile sensor-enabled devices along with the user comments to identify location specific advertisement opportunities. To best understand this technology, we discuss an example.

#### 3.1.1 Example 1

Daniel uses his smart phone to watch clips of his favourite CSI episodes while he was at a shopping centre waiting for his friend. The specific URL watched by Daniel is http://www.youtube.com/watch?v=sarYH0z948, which if one was to follow the link, will see that the video is shown without any advertisement. The video however attracted more than 5 million views, which has a significant viewership. Without advertisements, click-through opportunities are lost entirely. With random ones, the click-through success will be low. A significant opportunity that traditional television is unable to achieve in this case is to use analytical services from a sensor-cloud to identify monetisation keywords from the user comments. In our example, sun-glasses was mentioned. If this is combined with location information coming out of Daniels' smart phone, the analytical software could identify that a nearby shop is selling sun-glasses at a discount. It could then insert this advertisement in the video, which is likely to generate interest for Daniel to check out the sun-glasses shop.

#### 3.1.2 Research Problem

To achieve the targeted marketing outcomes presented in the example however, there are a number of research problems that has to be addressed on top of the sensor-cloud platform. First, the detection of monetisation keywords has to be made by considering the sentiment of a users comment. While sun-glasses is an opportunity to push an advertisement from a nearby optic shop, it may not be wise to do so if the comment about sun-glasses were negative, e.g., it makes the person looks ugly. As such, determining the keywords and knowing that it is appropriate to push for a sale is the first problem to be addressed here. Currently, sentiment analysis [9, 10, 11, 12, 13] do exists for text but in most cases, they work well only if the dictionary contains sufficient information about the words used in a comment, or that the training set is sufficiently large during the training phase of classification. Considering that YouTube comments are very different from traditional text, two things must be done before the above is possible. First, a dictionary for YouTube vocabulary must be built. Second, some automated (or semi-automated) mechanism is needed to assist in maintaining this dictionary to allow effective assessment of sentiment.

The other research problem is to identify monetisation opportunities by taking into account both the spatial and temporal information. Spatial information refers to location information such as the location of the user when the video is watched. This allows a spatial search to be conducted for advertisements that is relevant to the users locality. Temporal information refers to validity of an advertisement and the time in which a monetisation keyword (in a comment) was posted. In videos where comments are already posted for a while, one needs to consider if the comments still carry impact on the viewer. If not, then chances of a click-through is also likely to be low even though the opportunity is identified. Both spatial and temporal information is therefore as crucial as identifying the monetisation keywords. Gartner terms this context-aware computing, and has listed it as one of the top ten hot technologies in 2011 [14]. Clearly without doubt, the advertisement outcomes on a site like YouTube can be substantially improved if the three elements are properly identified. With a 100 million views a day, the sensor-cloud platform will carry the capacity and resources to enable such complex analytics.

### 3.2 Reporting

The next application that we want to discuss in this paper relates to the use of "*reporting*" as a model for knowledge discovery. We observed that traditional media now report news in a way that felt '*templated*'. A breaking news first appears which is often very brief at about two to three short paragraphs. In the next hour, the same article is expanded with additional information obtained from social sites. This brings us to question whether this process can be automated. Our initial investigation points to some degree of possibility, which we demonstrate by the following example using our third social application FaceBook.

FaceBook is a social networking site where the fundamental idea is to establish a personal profile and to link the profile to other profiles if a relationship exists. The relationship (or the link) becomes the basis for a user to be aware of what is happening with their friends and their activities. In many ways, it has similar functionalities such as YouTube videos or a Twitter tweet. In addition to a FaceBook version of video sharing and short messages FaceBook sports a very extensive photo album sharing mechanism. Often, it is easy to find users of FaceBook sharing all or some of their information to the public, i.e., people whom they do not know. This access to public information allows journalist to quickly write stories on traditional media, post links to photographs and videos that users on FaceBook produced. To some extend, this form of journalism is taking root at least on certain specific topics like product announcements, novelty stories and reports on public response relating to a major event, such as the earthquake example discussed above.

**3.2.1 Example 2**

Our current work explores a possible reporting model for product announcements. Consider the recent release of the Apple iPad 2. Being a major product announcements, we can easily obtain information about the thoughts of individuals, find out what they like and even gain links to their photos and videos discussing their out of the box experience. From various technological reports that we have seen on local newspapers, we note a structural reporting format that has the potential to be automated. Using two sources of information from FaceBook as our example, a search on FaceBook using their built in API allows us to find short messages posted on FaceBook regarding the keyword of interest. In this case, if we were to submit a search for iPad 2, we can get to a page such as http://www.facebook. com/pages/Ipad-2/135090849888347, which has many short messages, photos and videos about the product. Using different adaptors, which are small programs each with a specific role, we can quickly identify the information we need to form a news report. For ease of discussion, we list a few common adaptors that could be presented as a service within the sensor-cloud platform. This includes

- the identification of announcement information, e.g., timing, location, dates, etc.
- identify sentiment of an announcement, e.g., users are excited or hated the product, etc.
- identify reasons for a given sentiment class, e.g., what is it that users like about the new product.
- identify supplementary information, e.g., videos and photographs.

Each point above can be accomplished by an adaptor that could be plugged into a stream of search results such as represented by the URL given above. As it performs a very clearly defined and specific task, it is feasible in practice. The challenge of course is the need to use a wide range of technologies to develop individual adaptors and to have sufficient number of adaptors to reach a full article that matches the outputs of a journalist. Nevertheless, with just only the four 'adaptors' listed above, we can easily obtain the following paragraph using information from the results returned at http://www.facebook.com/pages/Ipad-2/135090849888347. The paragraph will read as follows, where normal texts represented the template and the bold texts are outputs from the adaptors listed above.

Apple has just announced that the iPad 2 will go on sale online at 1:00am on March 12, 2011. Responses seen on FaceBook is as expected with more than 14,000 positive 'likes' about the iPad 2. According to the 14,000+ FaceBook users, many look forward to the new video capture feature as well as the thinner form factor because it will now allow Apple fans to blog anywhere on the iPad. Initial photographs about the new iPad 2 can be seen at **http://www.facebook.com/lbum.php?fbid=144333632297402&id=135090849888347&aid=29911** and if a full review is of interest to readers, check out **http://www.facebook.com/appIeipad2?sk=app10442206389**.

**3.2.2 Research Problem**

The idea to achieve the above by humans is not difficult. Given an input consisting of text, photographs, videos and their interrelationships modeled as a graph, the human can easily provide answers to the adaptors listed. Doing this mechanically via algorithms however requires a lot more challenges. The first adaptor, which is to identify announcements can be seen as a problem that flows from our discussion in event detection. As such, the solution outline will not be too different from our discussion.

The second adaptor analyzes information to determine the overall sentiment. Sentiment analysis is a field of study that is currently popular amongWeb 2.0 sites. Notable examples of sites that are keen to use sentiment analysis includes sites like TripAdvisors and VirtualTourist so as to on-sell their business intelligence to hotel chains so that hotel chains can improve their customer satisfaction. Sentiment analysis are more mature in areas where the text is available. It is however weaker in learning

the true sentiment of a text that is made along with a photograph or video. With sites such as FaceBook generating large amount of data in different formats, how to determine sentiment as accurately as possible is a major challenge. While existing techniques presents a solution sketch, it is nevertheless insufficient. This is because current techniques focus almost solely on text while ignoring the social relationships of opinion posters as well as the temporal and spatial properties of social networking applications.

On the issue of just analyzing text, the length is also a problem. With most social sites supporting short messages, it is harder for traditional sentiment analysis techniques to work effectively. One possible solution is to build a dictionary to determine the sentiment of keywords. Once the keywords are determine, the individual sentiment is summed to obtain the net sentiment for the short sentence. We observe a different way to possible achieve the same result by finding a way to summarize multiple short text into a longer document [17] prior to sentiment analysis. This hasn't been done before and to our knowledge, there hasn't been discussion of why such a technique may fail. We believe it would be a worthy study no matter what the outcomes may be.

The third adaptor is to find the reasons for a given sentiment. This is in fact an extension of the sentiment problem. So far, sentiment analysis is concerned with one single outcome: is the sentiment positive or negative. Yet there hasn't been study on techniques that could explain why the sentiment is so. Hence, we felt that it is an important research issue and a direction that may be worth pursuing. Reasons for positive sentiments tell businesses what they have done right while the negative sentiments calls for reasons to understand where businesses have done wrong. We foresee natural language processing (NLP) to be a starting point. We also foresee techniques in summarization [17, 18] as another way to answer to this particular question.

In both cases, the general technique is to construct a graph of sentences and their relationships to identify which sentences are the important ones. The outputs of a summary is therefore the important sentences. We argue that when sentiments are negative or positive, the ones that lead the sentiment would also be the sentences that are likely to remain in the summary and thus, the summary will present answers to a given sentiment. Of course, there will be a need to extend and make modifications to the techniques. In particular, a lot of preprocessing will be needed given the large volume of social comments to be collected. Again, these low level preprocessing work can be moved to the sensor-cloud exposing such features as a software service in analytics, which is discussed in our framework in Section 2.

| Tweet | Topic Label Similarity | Score |
|---|---|---|
| the way that you flip your hair lolol i cant stand straight i need to melt you cause global warming | Anti Global Warming Documentary | 0.0301 |
| | Anti Global Warming Movie | 0.0298 |
| | Global Warming Lies | 0.0271 |
| global warming skeptics as knowledgeable about science as climate change believers, study says http://t.co/kukas6nz | Evidence Against Global Warming | 0.0836 |
| | Global Warming Skeptics | 0.0819 |
| | Global Warming Hoax | 0.0767 |
| #global-warming: global warming skeptics as knowledgeable about science as climate change ... - fox news http://t.co/ggfflmda | Evidence Against Global Warming | 0.0593 |
| | Anti Global Warming Movie | 0.0579 |
| | Global Warming Lies | 0.0574 |

Table 1. Search Query: Global Warming

**3.3 Topic Detection**

The last application that we want to focus on is the detection of topics in short texts. From our study of [21], we learned about how content is created by amateurs and shared within the community. The social sharing resulted in heavy interactions producing 'by products' such as short texts in the form of tweets, comments and social updates, etc. Frequently, they express the opinion of its authors. They can be about an event, a reference to an interesting URL, or a sentiment about a product (or service). While they can be easily dismissed on their own, short texts offer significant value (or impact) when they are looked at collectively. Researchers for example are studying the use of tweets to detect real-time events [22]. In the real-world, managing brand image on social sites by monitoring the texts posted online has become a serious business norm.

As short texts become a prevalent part of user generated content, a new form of information overloading is seen. Using the specific case of tweets as a generic representation of short texts, we can make the following statements. First, there are tweets that points to interesting content such as videos, photos, or Web pages. Then there are tweets that express sentiment about an

entity, product, or a service. And there are tweets that do not carry any "*business value*" or "*information value*". Additionally, mechanisms such as '*re-tweets*' heightens the amount of redundant information by having duplicates that are not immediately obvious. For example, re-tweets of the same URL can end up looking physically different. As a result, users can end up clicking on different URLs only to find themselves on the same page.

Clearly, the issue highlighted by the above is the potential for information overload. When large number of tweets are presented to a user, the ideal situation is to include a diverse set of tweets, exclude duplicates as well as those deemed '*worthless*'. One way to implement such a solution is to first determine the topic of a given tweet so that similar tweets can be grouped together to allow '*categories*' to be presented and subsequently, representative tweets from each category presented to deal with information overload. Clearly, a solution as such will extend to other forms of short texts and therefore, has wide applications. For example, the core algorithm of evaluating tweets could be applied to selecting a subset of comments for a YouTube video. This subset would represent the tens of thousands of comments that would otherwise take a long time to review. Similarly, the core algorithm could be applied to evaluating comments of Apps in the App store and in turn, the algorithm could be extended to create a better App search engine.

### 3.3.1 Example 3

Daniel is a university student working on a report about climate change. He is interested in incorporating some social opinions about the topic and therefore has turned to Twitter for some tweets related to the topic. David enters a search term such as '*global warming*' but gets tons of tweets that are related. Many of them however are duplicates, some saying the same thing in different ways and Daniel ended up scrolling through many pages of tweets to find a few good representations. Subsequently, Daniel found a new tool that provides a better search of social opinions. When the same search is entered, the results that returned are group by a given topic label. As seen in our experimental results, a search on global warming returns a set of tweets that were classified into topics such as '*global warming lies*', '*global warming sceptics*', '*global warming facts*', etc. The use of topic detection in his case has made it very easy for Daniel to see all the topics discussed on Twitter about '*global warming*' and he is able to look at a subset of tweets belonging to a topic. This makes the task of including social opinions in Daniel's report a lot easier as information overload gets addressed.

| Tweet | Topic Label Similarity | Score |
|---|---|---|
| sale #airfare #fly #alicesprings to #brisbane from $199 with qantas - http://t.co/ks7wm97a | Qantas Domestic Flights Australia | 0.0756 |
| | Qantas Domestic Flights | 0.0498 |
| | Qantas International Flights | 0.0480 |
| sale #airfare #fly #perth to #melbourne from $209 with qantas - http://t.co/1ruzsber | Qantas Flight Bookings | 0.0665 |
| | Qantas International Flights | 0.0651 |
| | Qantas US Flights | 0.0644 |
| sale #airfare #fly #sydney to #karratha from $319 with qantas - http://t.co/8yhhojg6 | Qantas Domestic Flights Australia | 0.0910 |
| | Qantas Flight Bookings | 0.0777 |
| | Qantas International Flights | 0.0769 |

Table 2. Search Query: Qantas

### 3.3.2 Research Problem

Topic detection is one of the methods to introduce diversity into the ranking of search results. Diversity ensures that results of a search do not end up with the same but instead, covers a broad range of concepts or sub-topics related to the search. From the research perspective of information retrieval, there are two ways of achieving diversity with search. The first achieves diversity on the basis of physical dissimilarity between the contents of each document in the search results [23, 25]. The second diversity mechanism works on ranking a diverse mix of sub-topics with respect to a given query [26, 24].

In the case of short texts like tweets, we note that physical similarity is hard to compute. In fact, we have conducted experiments on real tweets only to discover that the Jaccard distance [27] among tweets (from a given query) are either the same (i.e., a score of 1) or different (i.e., a score of 0). This bi-polar outcome, with a lack of variation in physical similarity, poses a problem for algorithms requiring similarity measures as the basis of operation.

The sub-topic label approach works on the basis that a given query will contain search results, where there exists sub-topics related to it. For example, a search for 'global warming' can be associated with sub-topics such as 'facts' or 'hoaxs', etc. ForWeb searches, determining the sub-topic is plausible as each search result points to a document with sufficient level of word terms to compute. When these methods are applied to tweets, the few terms within the 140 characters limit has made it hard to produce accurate results. This coincides with recent discussions that methods such as LDA [29] and classification [28], etc., performs poorly on short texts. As a result, neither diversity ranking approaches for Web documents work well on short texts.

In either approach, the challenge in dealing with short texts is due to their highly sparse representations through (i) the lack of sufficient word term co-occurrences, and (ii) the lack of context information in the short texts. To address this challenge, one of the ways is to select word terms in the short texts as a Web search query. This allows the short texts to be associated with a set of relevant Web documents and thus provides a logical way to augment the short text corpus with relevant word terms drawn from the Web documents. These augmented word terms then allow co-occurrences to be computed and context to be determined. This expansion of the short text corpus with relevant word terms is an approach broadly known as pseudo relevance feedback [30] within the information retrieval community. Another variation to this method is to match the short texts to topics learned from large knowledge repositories such as Wikipedia or the ODP [31, 32]. If a match is obtained, the text from the repositories or its topics are then merged with the short texts before they are processed by some text mining algorithms.

With good search results or where the knowledge repository is comprehensive, pseudo relevance feedback shows promising results. In situations where it works, the sub-topics discovered from short texts are indeed impressive. As reported in [33] where the right augmentation of word terms occurred, this method can even identify sub-topics that are not inherently evident in the original short text corpus. Such outcomes would have great application implications. However, pseudo relevance feedback is highly dependent on a good source of word terms augmentation. When we apply the top tweet and Web search terms to one such pseudo relevance feedback system, we find that the results turned out to be less ideal. In many cases, the system was unable to produce a right sub-topic label for a given tweet.

From our observation, the main issue is that the source of word terms augmentation is noisy and this is further complicated by inaccurate search matches. For example, when a search engine is used as the main source of word terms augmentation, the Web documents returned for a given tweet can contain advertisements, complex HTML formatting, irrelevant hyper links, etc., and such noisy information can cause unwanted terms to be augmented to the short texts. Consequently, the noise can affect the accuracy of sub-topic detection.

| Tweet | Topic Label Similarity | Score |
|---|---|---|
| @rubycairney: im so excited for the new season of pretty little liars :d | Pretty Little Liars Season | 20.0906 |
| | Pretty Little Liars | 0.0777 |
| | Pretty Little Liars Season | 30.0743 |
| @johnlemonnnn: next week is a pretty little liars monday! finally! | Pretty Little Liars Finale Date | 0.0698 |
| | Pretty Little Liars Last Episode | 0.0660 |
| | Pretty Little Liars Finale Video | 0.0608 |
| @nyikern: pretty little liars season 3 in 9 days #yup | Pretty Little Liars Series Finale | 0.0649 |
| | Pretty Little Liars Fans | 0.0647 |
| | Pretty Little Liars Finale Spoiler | 0.0622 |

Table 3. Search Query: Pretty Little Liars

Eddi's algorithm, TweeTopic, is an implementation of a pseudo relevance feedback system [33]. It consists of three main steps: (i) text transformation, (ii) search engine query, and (iii) text mining. From our attempts of developing and conducting evaluation of the Eddi's base algorithm – TweeTopic and subsequent variations developed to address the noise noted, we were able to draw a number of conclusions from the theory and the empirical results seen so far. Notably, we were able to drawn conclusions that Web documents that are noisy are not adequate for the reliable compute of its TF-IDF terms. A noisy Web document is hence one that exhibits one or more of the following characteristics.

- HTML that depends on JavaScripts, CSS styles to render the final content
- Flash content, or other formats of embedded objects in a HTML document
- Server-side scripting including forms, Ajax, etc.
- Heavily formatted sites through images or HTML layers, e.g., 
- Third party contents, e.g., advertisements that tries to match content and therefore acted as a decoy to the extraction algorithm.

and despite turning to different cleaning and content extraction tools, the results usually carries imperfections. Furthermore, the process is computationally heavy from our experience of our implementation due to the pre-processing and counting involved. While the response time looks acceptable for a single tweet, the solution is not scalable if huge volume of requests are made and the processing is to be carried out on the fly. It is possible to pre-compute many of the results but this would end up as a major operations that would be costly.

These problems translate into research questions on how then can we adequately achieve similar outcomes so that topic detection for short texts are as good as for Web documents. Our crucial observation is the level of noise present in modern Web documents. If no amounts of pre-processing the Web document can adequately help in achieving good results, then a solution that accepts the presence of noise is needed. Among these application ideas, we are pleased to report some initial positive results for topic detection. We present them in the next section.

**4. Experiment Results**

We note that the underlying concept of pseudo relevance feedback systems is great but not ideal when subjected to real-world operating conditions. In all our implementations of different algorithms, we continue to note that the results achieved are limited in terms of its impact. As the mode (e.g., presentation for desktop versus mobile devices) of information continues to evolve, with only more sophisticated HTML rendering, scripts, and other enhancements expected in the foreseeable future, the only conclusion we can drawn upon is that we will face more noisy documents than ever.

As long as noise is present, the pseudo relevance feedback for the short texts will not work precisely. Therefore, we felt that it is appropriate to rethink about how we could detect topic labels for our short text corpus. By disconnecting the relationship between the short text corpus and the Web document corpus represented by the Web, we were able to develop a more stable topic label detection system with consistently reasonable results. While this meant sacrificing some novel topics that could be potentially discovered, the overall performance of our proposed method [39], based on our initial experiments, strongly suggests the feasibility of our approach in practice.

| Tweet | Topic Label Similarity Score |
|---|---|
| two and a half men never gets old | Two Half Men CBS 0.1036 |
| | Two and Half Men TV 0.0963 |
| | Two Half Men Sheen 0.0869 |
| i'm watching two and a half men (33 others | 2.5 Men Ashton Kutcher 0.0733 |
| checked-in) http://t.co/ddvgf44p | Two Half Men Sheen 0.0707 |
| @getglue @twohalfmen_cbs | Two Men Ashton Kutcher 0.0689 |
| can't beat an episode of two and a | Two a Half Men Episodes 0.0932 |
| half men. #legendary Watch | Two Half Men Episodes 0.0683 |
| CBS | Two Half Men Show 0.0660 |

Table 4. Search Query: Two and a Half Men

Here, we present the results for four Twitter searches using the terms '*global warming*', '*qantas*', '*pretty little liars*' and '*two*

*and a half men*'. The first two terms are popular topics in Australia at the time when we conducted our experiments and the later two are taken from the list of top 50 search queries on Twitter in 2011. For each search, we show three tweets with each tweet being associated with the top three topic labels obtained via query prediction for normal Web search. Our method operates on the basis that query search and its related searches were in fact good proxies for topic labels. For example, when one searches for '*global warming*', the query prediction of a search engine will respond with suggestions such as '*global warming facts*', '*global warming sceptics*', '*global warming documents*', etc. We found that these query predictions are in fact the equivalent of human intervention in labelling sub-topics of a more generic query like '*global warming*'. More importantly, these query predictions were a result of heavy compute of the links users clicked on. These clicks are tracked and therefore, the search engine is able to learn the set of top relevant documents associated with each specific query. Using the top relevant documents, we were able to develop a feature space for each specific query prediction.

When a tweet is obtained, we performed the usual pseudo-relevance feedback process in the likes of [33]. The Web documents discovered for the tweet are then encoded into a feature space to describe the tweet. We then compare the feature space to each query prediction's feature space. The comparison is expressed as a Cosine similarity score. The higher the score, the higher the similarity of two feature spaces. And since we encode the entire Web document into a feature space, the noise in each document becomes insignificant and cancels one another during comparison. By comparing the different feature spaces associated with each query prediction suggestions, we were able to label the tweet accordingly to the query prediction. The results are shown in Tables 1 – 4 and for space reasons, we have only presented an overview here but we refer the reader to [39] for an in-depth discussion of our method and results.

We validated our results against a group of assessors. Each member in the group makes expert judgement on the topic label associated with the tweets and its associated topic labels. A scale of 1 to 10 was used to rate how relevant they think the topic label was for a given tweet. The inter-rater agreement calculated from their scores of forty tweets across the four topics proved to be significant and substantial. This was further confirmed by Flesis's Kappa [40] measure, which is a statistical measure for assessing how reliable the agreement between evaluators are. Such evaluations were appropriate for our tests as the scores are categorical in nature.

The outcomes of the human evaluation and the results from the Kappa measure can be explained by the fact that the topic labels drawn from query predictions are in fact a "*pseudo human labelling exercise*" when search engines monitors queries and the documents clicked. In other words, our method consequently ensures that a topic label makes sense since in retrospect, the query prediction is an outcome of a user's search refinement. Also, this means that our method can produce topic labels that are phrases compared to methods like LDA and TF-IDF, which are predominantly single word terms requiring further processing. For example, our method can directly produce phrases like '*frequent flyers*' instead of separate word terms '*frequent*' and '*flyers*'. This further translates to better assessment scores than topic labels capable of only single word terms.

## 5. Conclusion

In this paper, we presented a number of interesting applications that arise out of a platform like the sensor-cloud. Specifically, we presented a sensor-cloud from our earlier design as the basis to explore novel applications possible when data in the sensor-cloud are fussed with flexible computing resources and applications that utilizes the sensor-cloud platform as a service. For each problem, we presented associated research problems and discussed briefly how the sensor-cloud can enable such applications. This paper serves as the over-arching framework that will guide the research investigation undertaken as future works. Our discussion of the initial results obtained for the third application idea and the investigation reported [39] represents one of the realisation of the ideas discussed in this paper.

## References

[1] Akyildiz Ian, F.,Weilian Su., Yogesh Sankarasubramaniam., Erdal Cayirci, (2002). A Survey on Sensor Networks. IEEE *Communications Magazine*, 102-114 40 (8).

[2] Mohammad Mehedi Hassan, Biao Song., Eui-Nam Huh. (2009). A Framework of Sensor-Cloud Integration Opportunities and Challenges. *In*: Proceedings 3rd International Conference on Ubiquitous Information Management and Communication, 618-626.

[3] Bill Schilit, Norman Adams., RoyWant. (1994). Context-Aware Computing Applications. *In*: Proc.Workshop on Mobile Computing Systems and Applications, 85-90. IEEE Computer Society.

[4] Yu-Hsn Liu, Yongli Ren, Robert Dew, (2009). Monetising User Generated Content Using Data Mining Techniques. *In*: Proc. 8th Australiasian Data Mining Conference, 75-81, Melbourne, Australia.

[5] Meeyoung Cha., Haewoon Kwak., Pablo Rodriguez, Yong-Yeol Ahn, Sue Moon. (2007). I Tube, you Tube, everybody Tubes: Analyzing the Worlds Largest User Generated Content Video System. *In*: Proceedings 7th ACM SIGCOMM Internatonal Conference on Internet Measurement, 1-14, New York, NY, USA.

[6] Gary Becker., Richard Posner. (2011). The Future of Newspaper. http://www.becker-posner-blog.com/2009/06/the-future-of-newspapers-posner.html.

[7] Ali Moore. PBL Considers Further Media Sell-off. http://www.abc.net.au/lateline/business/items/200705/s1935762.htm.

[8] Hearst Marti, A. (1992). Direction-based Text Interpretation as an Information Access Refinement. Text-Based Intelligent Systems (Paul Jacobs, eds). Lawrence Erlbaum Associates.

[9] Das Sanjiv, R., Chen Mike, Y. (2007). Yahoo for Amazon! Sentiment Extraction from Small Talk on the Web. *Management Science*, 53 (9) 1375-1388.

[10] Tong Richard, M. An Operational System for Detecting and Tracking Opinions in on-line discussion. *In*: Proceedings SIGIR 2001 Workshop on Operational Text Classification in conj. in conjunction with ACM SIGIR New Orleans, USA.

[11] Turney Peter. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *In*: Proc. Association for Computational Linguistics 40th Anniversary Meeting (Pierre Isabelle, eds), 417-424, Philadelphia, PA, USA.

[12] Xiaowen Ding., Bing Liu., Lei Zhang. (2009). Entity Discovery and Assignment for Opinion Mining Applications. *In*: Proceedings 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 1125-1134.

[13] Weining Wang., Yingling Yu., Jianchao Zhang. (2005). A New SVM Based Emotional Classification of Images, *Journal of Electronics*, 22 (1) 98-104, Springer.

[14] Susan Moore. (2011). Gartner Says Context-Aware Computing Will Be a $12 Billion Market By 2012. http://www.gartner.com/it/page.jsp?id=1229413.

[15] Stacey Higginbotham. (2011). ensor Networks Top Social Networks for Big Data, Bloomberg BusinessWeek,http://www.businessweek.com/technology/content/sep2010/tc20100914_284956.htm.

[16] Adam Ostrow. (2011)Japan Earthquake Shakes Twitter Users. And Beyonce. http://mashable.com/2009/08/12/japan-earthquake/.

[17] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. (2000). Multi-Document Summarization by Sentence Extraction. *In*: Proceedings 2000 NAACL-ANLPWorkshop on Automatic Summarizatio in conj. Association for Computational Linguistics, 40-48, Stroudsburg, USA.

[18] Sun Park, Ju-Hong Lee, Chan-Min Ahn, Jun Sik Hong., Seok-Ju Chun. (2007). Multi-document Summarization Based on Cluster Using Non-negative Matrix Factorization. *In*: Proceedings of Lecture Notes in Computer Science, 761-770, 4362.

[19] Yuzhang Han., Ivan Janciak., Peter Brezany., Andrzej Goscinski. (2011). The CloudMiner - Moving Data Mining into Computational Clouds, in Grid and Cloud Database Management (G. Aloisio and S. Fiore, eds), 193-214.

[20] Caroline McCarthy. (2011). Nielsen: Twitter's Growing Really, Really, Really, Really Fast. CNet News. http://news.cnet.com/8301-13577_3-10200161-36.html.

[21] Yu-Hsn Liu, Kok-Leong Ong , Andrzej Goscinski. (2012). Sensor-Cloud Computing: Novel Applications and Research Problems. Fourth International Conference on Networked Digital Technologies, Dubai, India.

[22] Takeshi Sakaki., Makoto Okazaki., Yutaka Matsuo. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *In*: Proceedings of the 19th International Conference onWorldWideWeb, 851-860.

[23] Jaime Carbonell., Jade Goldstein. (1998). The Use of MMR, Diversity-Based Rranking for Reordering Documents and Producing Summaries. *In*: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 335-336.

[24] Clarke Charles, L. A., Maheedhar Kolla., Cormack Gordon., V., Olga Vechtomova., Azin Ashkan., Stefan Buttcher., Ian MacKinnon. (2008). Novelty and Diversity in Information Retrieval Evaluation. *In*: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Rtrieval, 659- 666.

[25] Xiaojin Zhu., Andrew, B. (2007). Goldberg and Jurgen Van and and Gael David Andrzejewski. Improving Diversity in Ranking Using Absorbing RandomWalks. *In*: Physics Laboratory University of Washington, 97-104.

[26] Cheng Xiang Zhai, Cohen William, W., John Lafferty. (2003). Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. *In*: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and development in Informaion Retrieval, 10-17.

[27] Gower John, C. (2005). Similarity, Dissimilarity and Distance Measure. John Wiley & Sons, Ltd.

[28] Manning Christopher, D., Prabhakar Raghavan., Hinrich Schtze. (2008). Introduction to Information Retrieval. Cambridge University Press, New York.

[29] Blei David, M., Ng Andrew, Y., Jordan Michael. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res*., 3 993- 1022.

[30] Ou Jin., Liu Nathan, N., Kai Zhao, Yong Yu, Qiang Yang. (2011). Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering. *In*: Proceedings of the 20th ACM International Conference on Information and knowledge Management, 775-784.

[31] Xuan-Hieu Phan, Le-Minh Nguyen, Susumu Horiguchi. (2008). Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections. *In*: Proceedings of the 17th International Conference on World Wide Web, 91-100.

[32] Xuan-Hieu Phan., Cam-Tu Nguyen., Dieu-Thu Le., Le-Minh Nguyen., Susumu Horiguchi., Quang-Thuy Ha. (2011). A Hidden Topic-Based Framework Toward Building Applications With Short Web Documents, *IEEE Trans. on Knowl. and Data Eng.*, 23 (7) 961-976.

[33] Bernstein Michael, S., Bongwon Suh., Lichan Hong., Jilin Chen., Sanjay Kairam., Ed H. Chi. Eddi. (2010). Interactive Topic-Based Browsing of Social Status Streams. *In*: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, 303-312.

[34] Michael Bendersky., Bruce Croft, W. (2003). iscovering key concepts in verbose queries. *In*: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 491-498.

[35] Anette Hulth. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *In*: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 216-223, Stroudsburg, PA, USA. Association for Computational Linguistics.

[36] Kristina Toutanova and Christopher Manning. (2000). Enriching The Knowledge Sources Used *In*:A Maximum Entropy Part-of-Speech Tagger. *In*: Joint SIG- DAT Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-2000).

[37] Ruben Rios and Javier Lopez. (2011). Exploiting Context-Awareness to Enhance Source-Location Privacy in Wireless Sensor Networks. The Computer Journal, 1603-1615, 54, Oxford University Press.

[38] Rosi, A., Mamei, M., Zambonelli, F., Dobson, S., Stevenson, S., Juan Ye. (2011). Social Sensors and Pervasive Services: Approaches and Perspectives. Pervasive Computing and CommunicationsWorkshops (PERCOMWorkshops), 2011 IEEE International Conference on 525-530, 21-25.

[39] Jing Zhang, Yu-Hsn Liu., Kok-Leong Ong. (2012). Diversity Ranking Algorithms for Short Texts. Technical Report (TR 12/11). School of Information Technology, Deakin University.

[40] Fleiss, J.L., et al. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76 (5) 378 –382.