Towards a Visually Enhanced Medical Search Engine

Lavish Lalwani[1,2], Guido Zuccon[1], Mohamed Sharaf[2], Anthony Nguyen[1]

[1] The Australian e-Health Research Centre, Brisbane, Queensland, Australia;
[2] The University of Queensland, Brisbane, Queensland, Australia.
lavish.lalwani@uqconnect.edu.au, m.sharaf@uq.edu.au,
{guido.zuccon, anthony.nguyen}@csiro.au

Abstract. This paper presents the prototype of an information retrieval system for medical records that utilises visualisation techniques, namely word clouds and timelines. The system simplifies and assists information seeking tasks within the medical domain. Access to patient medical information can be time consuming as it requires practitioners to review a large number of electronic medical records to find relevant information. Presenting a summary of the content of a medical document by means of a word cloud may permit information seekers to decide upon the relevance of a document to their information need in a simple and time-effective manner. We extend this intuition, by mapping word clouds of electronic medical records onto a timeline, to provide temporal information to the user. This allows exploring word clouds in the context of a patient's medical history. To enhance the presentation of word clouds, we also provide the means for calculating aggregations and differences between patient's word clouds.

Keywords. Visualisation, Timeline, Word Cloud, Medical Search.

Introduction

Current information systems deployed in clinical settings require practitioners and information seekers to review all medical records for a patient or enter database-like queries in order to retrieve patient information. Clinical data is often organised primarily by data source, without supporting the cognitive information seeking processes of clinicians and other possible users. For example, "The Viewer" application deployed by Queensland Health allows clinicians to access all patient electronic medical records collected by Queensland Health hospitals and facilities[1]. To access this information, clinicians need to enter data that allows them to select a patient (e.g., name, date of birth, Medicare number, etc.); afterwards they are given access to all information collected for that patient. However, they are unable to search through the medical records of the selected patient: if clinicians require a patient's past medical history, they have to read all medical records for that patient (organised by type of data, e.g. discharge notes, laboratory reports, etc., and clinical facility). This can be a very time consuming and tedious way of accessing information, particularly when clinicians

[1] Electronic medical record viewer solution, http://www.health.qld.gov.au/ehealth/theviewer.asp

want to review a large number of cases for research purposes, e.g. observe the effect a treatment had on their patient population.

An alternative solution is to deploy an information retrieval system where searches over patient records can be conducted with keywords, and medical records are ranked against the user query. We argue that this is a more efficient way for accessing patient information; previous research has developed systems that are able to search for relevant information in medical records [1, 2]. This paper considers how these systems could be improved by enhancing the presentation of results retrieved in answer to information seekers' queries. Search results are commonly shown to users as textual snippets that attempt to capture relevant portions of the medical record. Since these snippets are small chunks of text extracted from the original document (extractive summarisation), they often lack important information or can be misleading, especially if the original document is a medical record [3]. In addition, textual snippets do not convey an overview of the general clinical picture of a patient. For this reason, it is difficult to determine whether a medical case matches a search and whether it should be explored further; this thus requires the information seeker to access and read much of the document to determine its relevance to the query.

This paper investigates the use of data visualisation as a means for solving this problem. Data visualisation has the potential to provide a meaningful overview of medical reports, visits or even a patient's life and therefore may assist searchers to determine whether a medical document is relevant and worth further examination. Data visualisation may provide a simpler approach to augment standard searching methods for medical data. The remainder of the paper describes a system prototype that implements two data visualisation techniques: word clouds and timelines.

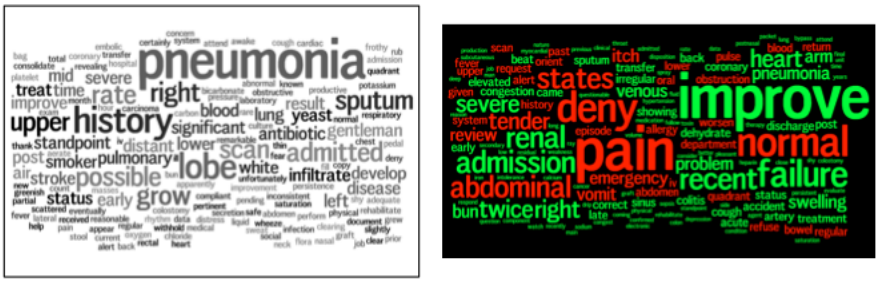
1. Related Work

Word clouds provide a visual representation of the content of a document by displaying words considered important in a document. Words are arranged to form a cloud of words of different sizes. The size of a word within a cloud is used to represent the importance of that word in the document; often, the importance of a word is computed as a function of the frequency of that word within a document. Figure 1 shows examples of word clouds.

In this paper we posit that word clouds have the ability to provide a better summary of the information contained in a medical record than textual snippets. This is supported by existing research on employing word clouds within information retrieval systems. For example, Gottron used a technique akin to word clouds to present news web pages [4]. In that study it was found that word clouds helped users to decide upon the relevance of news articles to their search query. Kaptein and Marx used word clouds to enhance information access to debate transcripts from the Dutch parliament [5]; they found that word clouds provided an effective first impression of the content of a debate.

Timelines are an additional data visualisation technique providing a map of events over time. The visualisation of events on a timeline provides the user with information related to which events occurred prior (and after) to an event of interest;. In our scenario, medical records belonging to a patient represents an event. Visualising medical records over a timeline allows for the possibility of mapping an entire patient's medical history within a unique visual representation. Previous research found that

employing timelines for displaying patient medical records has the benefit of enabling clinical audit, reduced clinical errors, and improved patient safety [6]. Bui et al. have explored the use of timelines to give a problem-centric visualisation of medical reports, where patient reports are organised around diseases and conditions and mapped to a timeline [7].



(a) An example of word cloud computed from a medical record. (b) An example of differential word cloud.

Figure 1. Word clouds computed from a medical record.

2. Word Clouds and Timelines

As supported by the previous research already outlined, this paper posits that word clouds and timelines can be effective visualisation techniques to provide quick information access to clinical records. The clinical records used to develop the prototype system were obtained from the TREC Medical Records Track corpus, a collection of 100,866 medical record documents taken from U.S. hospitals. Note that documents belonging to a single patient's admission were grouped together, obtaining a total of 17,198 groups of records. Next, we present the algorithms used within the system to generate word clouds and timelines.

2.1. Word Cloud Generation

The generation of a word cloud within our prototype system is a multi-step process.

The first step consists of removing tokens and words from the documents that convey limited or no information (stop word removal). These may include symbols, special characters, and words contained in a 'stoplist' (e.g. "the", "a", "when", etc.). This step is used to avoid displaying irrelevant or non-informational words within the word clouds.

The second step involves stemming the text of the medical reports. Stemming consists of reducing a word to its base form (stem). Stemming is applied to conflate syntactical variations of the same word (e.g. plurals, gerund forms, past tense, etc.) into a single token to represent the fact that they may have the same or similar meaning.

The third step consists of generating a probability distribution over the vocabulary words w, in a document d, $P(w|d)$. Since a word cloud cannot display all the words in a document, this distribution is used to derive the list of words that will form the word cloud and their final font size (step four). Language models are used to compute such probability distributions. The probability of a word w in a document d is computed as a

function of the occurrence of w in the medical records as the following equation mathematically explains.

$$P_{\lambda}(w \mid d) = (1-\lambda)P(w \mid d) + \lambda P(w \mid C) \quad (1)$$

In Equation 1, $P(w|d)$ is calculated as the ratio between the number of occurrences of w in d and the total number of words in d (maximum likelihood estimate). Similarly, $P(w|C)$ is calculated as the ratio between the number of occurrences of w in the whole corpus of medical reports C and the total number of words in C. These probabilities are interpolated according to the parameter λ, which controls the importance of background information (i.e., $P(w|C)$) when determining the importance of word w in the context of document d. The use of both the maximum likelihood estimate and the background language modelling are referred to as Jelinek-Mercer smoothing; more details on language modelling can be found in [8].

The last step (fourth step) is the generation of the actual word cloud. Words in a document are ranked in decreasing order of their probability $P(w|d)$, and only the top ranked words are selected to be included in the word cloud. The probabilities of the selected words are mapped into font sizes, and the appropriately sized words are placed in the word cloud for document d. Figure 1a shows an example of a word cloud generated from a patient medical report.

2.2. Word Cloud Aggregation

Individual word clouds could be merged to visualise an entire patient hospital visit or medical history as a unique word cloud. Two word clouds wc_1 and wc_2 are merged according to the following equation:

$$P(w) = P(w \mid wc_1)P(wc_1) + P(w \mid wc_2)P(wc_2) \quad (2)$$

where $P(w|wc_i)$ represents the probability[2] of word w in word cloud wc_i, and $P(wc_i)$ is the probability associated to wc_i. Currently, we consider word clouds to be uniformly distributed (thus $P(wc_1) = P(wc_2)$); however future developments may consider biasing word clouds according to temporal relations or document types when merging. As previously stated, Equation 2 can also be used to create a word cloud representing a complete patient medical history by merging all the word clouds associated to their medical records. Similarly, Equation 2 can be applied for merging word clouds associated with reports belonging to different patients.

2.3. Word Cloud Differential

A differential word cloud is designed to highlight the differences between two word clouds (i.e. between two documents). Since two word clouds are effectively two probability distributions, their difference can be computed using the Kullback-Leibler (KL) divergence. Equation 3 provides the means for computing the difference between word clouds, given the source word clouds wc_1 and wc_2.

[2] $P(w|wc_i)$ is equivalent to $P(w|d)$ if wc_i represents the word cloud for document d; however, note that wc_i may have also be computed from the merging of other previously computed word clouds.

$$D_{KL}(wc_1 \,||\, wc_2) = \sum_i P(w_i \mid wc_1) \log \frac{P(w_i \mid wc_1)}{P(w_i \mid wc_2)} \quad (3)$$

The magnitude of the KL divergence can be thought of as the degree of difference between the two word clouds. The value of KL divergence for each word can be used to generate a word cloud that provides visual information about how the two original word clouds differ. We refer to this type of word cloud as a differential word cloud (between wc_1 and wc_2). In a differential word cloud, the sign of D_{KL} for each word (i.e. $D_{KL}(w,wc_1||w,wc_2) = P(w|wc_1)log[P(w|wc_1)/P(w|wc_2)]$) determines the colour the word would be painted with. Words with positive D_{KL} values are painted green and words with a negative D_{KL} values are painted red. In this case, if a word is painted green it means it has a stronger presence (i.e. higher probability) in wc_1. The degree to which this presence is stronger is signified by the size of the word in the cloud (the bigger the word, the stronger the difference in presence). The opposite applies for a red colour word in the differential word cloud. Note that if the calculation was conducted with the probabilities in reverse order, the colours on the differential word cloud will reverse. An example of a differential word cloud is shown in Figure 1b.

2.4. Timeline Generation

The generation of timelines involves, for each medical report, extracting the date and time it was created. This was achieved using metadata information present in the reports from the TREC Medical Records Track corpus; however, it is acceptable to assume that similar metadata is present in records from other hospital providers. Since entire patient admissions were mapped to timelines, after dates and times are extracted for all records in a patient admission, this metadata, along with the medical record data are rendered within a timeline created using the Java Script library, Timeline JS[3]. This means that when retrieving a particular medical record, it can be displayed within context of the other reports produced for that patient admission.

3. Integration of Word Clouds and Timelines

The prototype described here is a modular information retrieval system, developed based on the Apache Lucene 4 framework, specifically for searching archives of medical records. Its architecture consists of three main modules: the indexer, the visualiser, and the searcher.

Within the indexer module, medical records are parsed and stored within a representation appropriate for supporting the retrieval stage (inverted file). The indexer is built using the Apache Lucene 4.0 incremental indexing capabilities, thus allowing new documents to be included in the index without re-indexing the previous documents. The indexer also maintains the relation between medical records and patients.

The searcher module is responsible for retrieving documents from the index that match a user query. A ranked list of medical admissions is produced as the result of querying the system.

[3] http://timeline.verite.co/

The visualiser module has the responsibility of rendering the results of a search and supporting navigation across search results. The modular architecture of the system integrates the visualisation methods described in Section 2 within the visualiser module without modifying the approaches used to index and retrieve documents. Indeed, the visualiser module is independent of the processes used in the other modules, allowing for flexibility when devising and testing new visualisation algorithms, as well as deploying versions of the system tailored to specific scenarios. Figure 2 shows a screenshot of an implementation of the methods described in Section 2 within the prototype system visualiser module. The figure illustrated a situation where a user has submitted a query and is in the process of examining a specific medical record. The content of the record is rendered as a word cloud allowing the user to quickly understand the content of the record itself. The text of the record can be accessed through the "Reports view" button above the word cloud. The record is also placed within the timeline of the patient admission to the hospital (bottom of Figure 2).

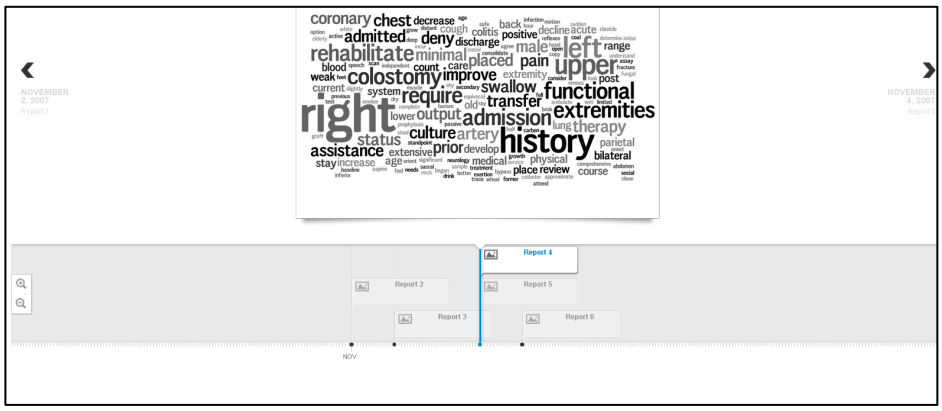


Figure 2. A screenshot of the visual interface of the system showing the use of word clouds and timelines.

4. Conclusion

In this paper we have presented two techniques, word clouds and timelines, to enhance search results presentation within medical records search. Word clouds have the potential to provide a rapid overview of an entire medical report, admission and patient history. Timelines provide a visual means to represent patient journeys as well as to place a medical record within the temporal context of other existing records. These techniques were integrated within the visualiser module of our prototype, a state-of-the-art medical information retrieval system. Future work will be directed towards a formal evaluation of the proposed techniques in a real scenario. Possible improvements will consider n-grams (sequences of *n* words, e.g. 'heart attack') and medical concept detection and reasoning (e.g. "heart attack" and "myocardial infarction" within a record should contribute towards the same medical concept) when building and rendering word clouds.

References

[1] Voorhees, E., & Tong, R. Overview of the TREC 2011 Medical Records Track. *In Proceedings of TREC* (2011).
[2] Zuccon, G., Koopman, B., Nguyen, A., Vickers, D., & Butt, L. Exploiting Medical Hierarchies for Concept-Based Information Retrieval. *In Proceedings of ADCS* (2012), 111-114.
[3] S. Afantenos, V. Karkaletsis & P. Stamatopoulos, Summarization from Medical Documents: a survey, *Artificial Intelligence in Medicine* **33** (2005), 157-177.
[4] T. Gottron, Document Word Clouds: Visualising Web Documents as Tag Clouds to Aid Users in Relevance Decisions, *Lecture Notes in Computer Science* **5714** (2009), 94-105.
[5] Kaptein, Rianne, and Maarten Marx. Focused Retrieval and Result Aggregation with Political Data. *Information retrieval* 13.5 (2010): 412-433.
[6] Gill, J., Chearman, T., Carey, M., Nijjer, S., & Cross, F. Presenting Patient Data in the Electronic Care Record: the role of timelines. JRSM short reports, 1(4), (2010).
[7] Bui, A. A., Aberle, D. R., & Kangarloo, H. TimeLine: visualizing integrated patient records. Information Technology in Biomedicine, IEEE Transactions on, 11(4), (2007), 462-473.
[8] Zhai, C. Statistical Language Models for Information Retrieval. Synthesis Lectures on Human Language Technologies, 1(1), (2008), 1-141.