Topic Modelling of Patient Opinion

A minor thesis submitted in partial fulfilment of the requirements for the degree of

Masters of Computer Science

Bin Lu

School of Computer Science and Information Technology

Science, Engineering, and Technology Portfolio,

Royal Melbourne Institute of Technology

Melbourne, Victoria, Australia

October 14, 2014

Declaration

This thesis contains work that has not been submitted previously, in whole or in part, for any other academic award and is solely my original research, except where acknowledged.

This work has been carried out since TODO:MONTH TODO:YEAR, under the supervision of Dr Jenny Zhang, Dr Amanda Kimpton, Dr Daryl D'Souza.

Bin Lu

School of Computer Science and Information Technology

Royal Melbourne Institute of Technology

October 14, 2014

Acknowledgements

TODO:THANKS!

Contents

List of Figures

List of Tables

Abstract

Chapter 1

Introduction

Publicly available opinions and service feedback provide valuable informations for decision making for both service providers and consumers. With the help of websites, blogs, forums and social networks, it is never been so easy to express opinions and leave feedback. Analysing the opinions becomes a challenge, not just because of the quantity of the data, most opinion from general users are free form text. The massive quantity of the data wont be effectively used until there is a systematically approach of analysing and summarizing. Many techniques have been proposed to solve this problem. MDK-LDA model proposed by Chen(Chen et al. [2013]) , the method extends the Latent Dirichlet Allocation(Blei et al. [2003]), the later one becoming the standard method in topic modelling and been extended in variety ways. The basic idea of LDA is treat each document in a collection as a vector of word count, each document is represented as a probability distribution over a number of topics, while each topic is represented as a probability distribution over a number of words. MDK-LDA introduces a new latent variable s in LDA to model s-sets. Each document is an admixture of latent topics while each topic is a probability distribution over s-sets. Another approach is Aspect-based

Summarization(Garcia-Moya.L and Berlanga-Llavori.R [2013]), it is usually composed of three main tasks: aspect identification, sentiment classification, and aspect rating. Generally this model is used to analysing product review, it is designed to effectively retrieve features and sentiment for products.

Most previous studies focus on analysing product reviews. We are interested to discover some model that suite service reviews. More specifically, reviews relate to healthcare. Study shows the effective governance is increasingly recognized as pivotal to improvements in healthcare quality(Bismark and Studdert [2013]), moreover current issue of effectiveness of the authority is affected by insufficient resource and inadequate information received(Bismark et al. [2013]). The object we are going to study is www.patientopinion.org.au, it is a publicly available healthcare forum. It allows user to post their own healthcare related story, the stories are not restricted from patient, it can also from hospital workers, nurses or doctors. The story can be positive or negative or a bit from both side. Although the story body is free form text, user still has to follow a certain template while submit the story.

Due to the unique characteristic of the data from Patient Opinion, the existing models of topic modelling may not give the best result, on other hand LDA has been approved a very effective model, and been used as a based model in many topic modelling studies. We choose LDA as our base model, and incorporate unique feature in Patient Opinion, specifically the section of Whats Good and What could be improved. These two sections are filled in by user while submitting the story, the template is provided by the website. Generally this will be the main topic or features user want to give feedback about in the story. And we assume user labelled story 100% accurate. The question we aim to answer in this thesis:

- How to use user specified features to improve the performance and accuracy in topic

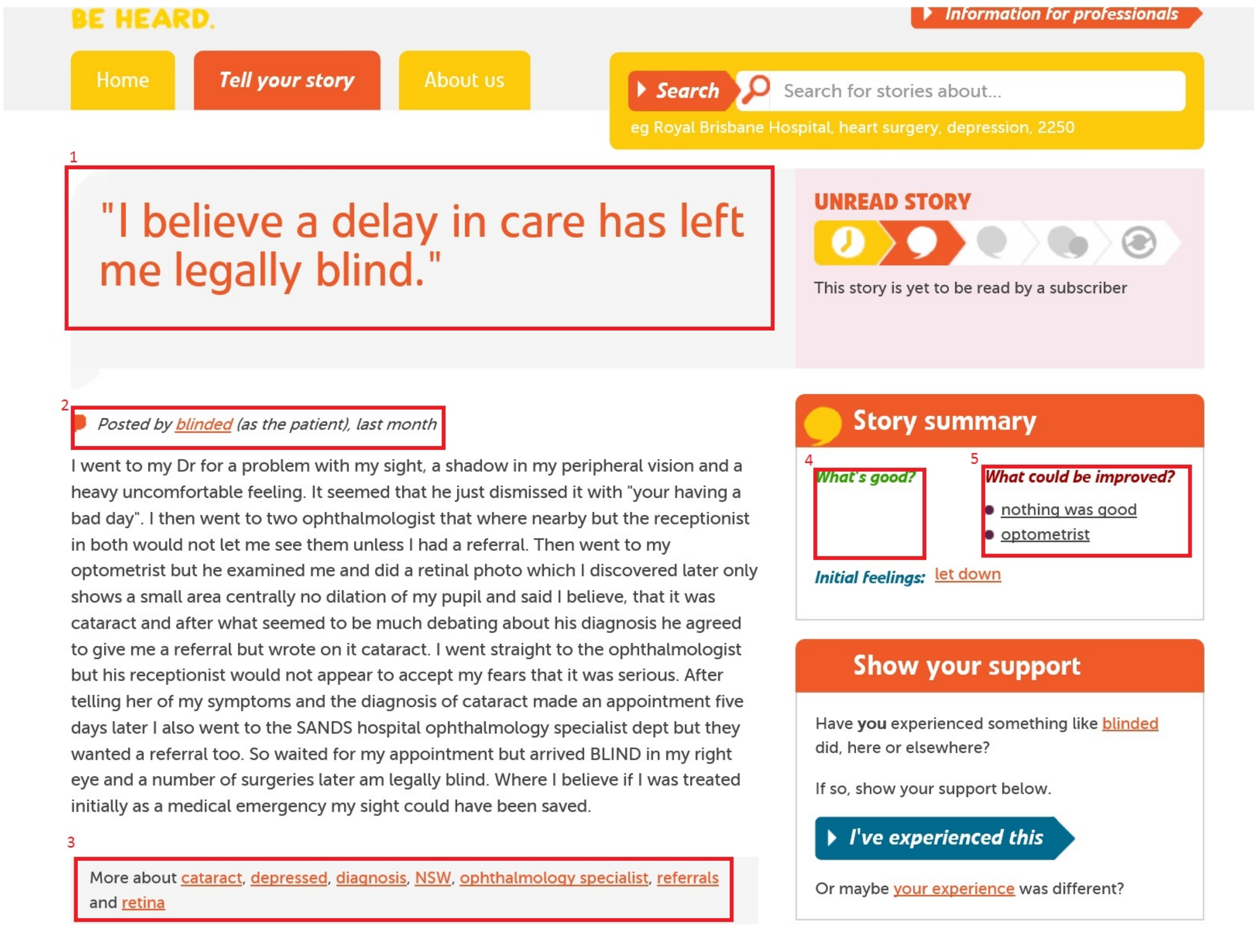

Figure 1.1: Patient Opinion Story Sample

```
    <article id="story" data-po-opinionid="59518" itemscope itemtype="http://data-vocabulary.org/Review">

<h1>
    <span class="top_dec"></span>
    <blockquote>
                                                                                                1
        &quot;<span id="opinion_title" itemprop="summary" class="">I believe a delay in care has left me legally blind.</span>&quot;
    </blockquote>
    <span class="btm_dec"></span>
</h1>

<p class="info">

         Posted by
                <span itemprop="reviewer"><a href="/opinions?author=blinded" title="Other opinions from blinded">blinded</a></span>
                                                                                2
        (as <span id="opinion_author_role" class="">the patient</span>),
    <time itemprop="dtreviewed" datetime="2014-07-22T04:35:56Z" title="Submitted on 22/07/2014 at 04:35 and published by Patient Opinion on 04/08/2014 at 05:04">last month</time>
</p>

    <div class="story_copy">
        <blockquote id="opinion_body" itemprop="description" class="text ">
            <p>I went to my Dr for a problem with my sight, a shadow in my peripheral vision and a heavy uncomfortable feeling.  It seemed that he just dismissed it with "your having a bad day".  I then went to two ophthalmologist that where nearby but the receptionist in both would not let me see them unless I had a referral.  Then went to my optometrist but he examined me and did a retinal photo which I discovered later only shows a small area centrally no dilation of my pupil and said I believe, that it was cataract and after what seemed to be much debating about his diagnosis he agreed to give me a referral but wrote on it cataract.  I went straight to the ophthalmologist but his receptionist would not appear to accept my fears that it was serious.  After telling her of my symptoms and the diagnosis of cataract made an appointment five days later I also went to the SANDS hospital ophthalmology specialist dept but they wanted a referral too. So waited for my appointment but arrived BLIND in my right eye and a number of surgeries later am legally blind.  Where I believe if I was treated initially as a medical emergency my sight could have been saved.</p>
        </blockquote>

    </div>

    <div class="related clearfix">
        <p> 3
            More about <a href="/opinions/tags/cataract">cataract</a>, <a href="/opinions/tags/depressed">depressed</a>, <a href="/opinions/tags/diagnosis">diagnosis</a>, <a href="/opinions/tags/nsw">NSW</a>, <a href="/opinions/tags/ophthalmology%20specialist">ophthalmology specialist</a>, <a href="/opinions/tags/referrals">referrals</a> and <a href="/opinions/tags/retina">retina</a>
        </p>
    </div>
```

Figure 1.2: Patient Opinion Story Sample Source 1

modelling.

- What is the distribution of topics over locations (State level).

```
<div class="module standard_module" id="saying">
    <h2>
        Story summary</h2>
    <div class="inner">
        <ul class="left">
            <h3 class="green">What's good?</h3>

        </ul>
        <ul class="right">
            <h3 class="red">What could be improved?</h3>

                <li><a href="/opinions?tag=nothing%20was%20good">nothing was good</a></li>
                <li><a href="/opinions?tag=optometrist">optometrist</a></li>
        </ul>

            <ul class="lower">
                <h3 class="blue">Initial feelings:</h3>

                <a href="/opinions/tags/let%20down">let down</a>
            </ul>
    </div>
</div>
```

Figure 1.3: Patient Opinion Story Sample Source 2

Chapter 2

Related Works

2.1 LDA

Also known as Latent Dirichlet Allocation or discrete PCA is a Bayesian graphical model for text document collections represented by bags-of-words (Newman et al. [2009], Blei et al. [2003], Griffiths and Steyvers [2004], Buntine and Jakulin [2004]). The model allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Generally, only a small number of words have high likelihood in each topic and each document only presents certain number of topics. Following is the equation of collapsed Gibbs sampling:

$$p(z_{id} = t \mid x_{id} = w, Z^{\neg id})\alpha \tag{2.1}$$

$$\frac{N_{wt}^{\neg id} + \beta}{\sum_w N_{wt}^{\neg id} + W\beta} \frac{N_{wt}^{\neg id} + \alpha}{\sum_w N_{wt}^{\neg id} + T\alpha} \tag{2.2}$$

2.2 MDK-LDA

As mentioned before, LDA is a powerful topic modeling framework, however recent studies found that these unsupervised models may not produce topics that conform to the user's existing knowledge(Chen et al. [2013]). Chen et al (Chen et al. [2013]) proposed a novel knowledge-based model, called MDK-LDA, which is capable of using prior knowledge from multiple domains to help topic modeling in the new domain. A new latent variable 's' is added to model the s-set, each document represent admixture of latent topics while each topic is a probability distribution over s-set. MDK-LDA uses s-set to distinguish topics in multiple senses. For example the world light can be represented by two s-set: light, heavy, weight and light, bright, luminance, if light co-occurs with bright or luminance it will be assigned to s-set light, bright, luminance.

Chapter 3

The Approach

Although LDA provides a powerful framework for extracting latent topics in text document, but sometimes learned topics are lists of words that do not convey much useful information (Newman et al. [2009]). Some extrinsic evaluation has been used to demonstrate the effectiveness of the learned topic in the application domain, but standardly, no attempt has been made to perform intrinsic evaluation of the topics themselves, either qualitatively or quantitatively (Newman et al. [2010]). To solve the problem, base LDA model had been extended either by incorporating human judgement in to the model-learning framework or creating a computational proxy that simulates human judgements (Chang et al. [2009]), for example the MDK-LDA model (Chen et al. [2013]) we introduced in section 2. Due to the unique characteristic of the data of Patient Opinion, we use user input to simulate human judgement, hence to produce a better quality topic modelling result.

3.1 Description of Data

Data from Patient Opinion contains many informations, however we only interested in few parts of them in our project Figure 1: 1) The title of the story, its the summary of story by the user. 2) The author role and time of the post, the role could be the patient, patients relative, carer or doctor. 3) The more about section is from website moderator, it inserts relevant tags to the story. 4) & 5) are the most important fields to our project, these field are inserted by the user, the fields indicate what user thinks the story is about, and we use these fields to simulate user judgement in topic modelling.

3.2 Preprocessing of Data

Everything been converted to lower-case, collect all unique words in user specified field. This collection is used to filter out the words in each topic that generated by LDA. A list of related document ID to each word also collected, see Figure3.1.

3.3 Using user input to improve topic modelling result

Topics learned from LDA sometimes dont convey much useful information, sometime it is caused by overfeeding the result set, for example it will include top 20 words for each topic (based on the settings, the total number in each topic can be configured), some words may not make any sense in current topic but statistically significant to the topic. Our goal is try to use user input to reduce the noise while retaining as much information as possible to describe or label the topic. The generative process is given as follows:

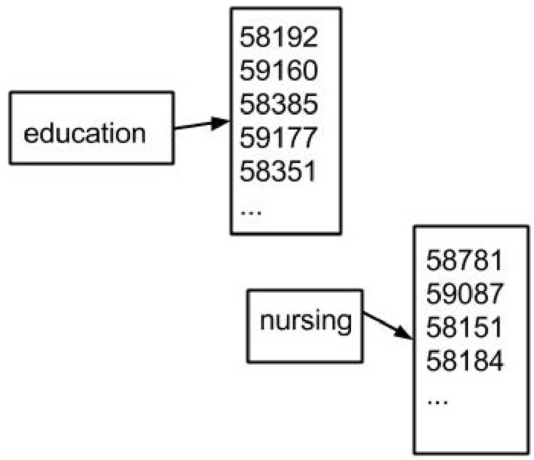

Figure 3.1: Patient Opinion Story Sample

1. Collect unique words from user specified field as S set.

2. Generate a set of topics T with LDA model.

3. Calculate result set R as: for each topic $t \in \{1, ..., T\}, r_n = t_n \cap S$

Chapter 4

Experiments and Result

We collected all 624 stories from Patient Opinion by August 2014. The count of unique user specified term is 659. 100 topics are generated using Mallet [1] with setting of optimize interval equals to 20. Table4.1 shows top 12 topics, it is ranked by the number of terms in original topic that matches the user specified terms

The total number of terms in the result set R is 527, compare to 2000 in original T set. Table4.1 shows the sum of composition for each term in the topic in set R and T

4.1 Topic Coherence Evaluation

Apart from quantitative and qualitative evaluation as above, evaluating topic coherence is a component of the larger question of what are good topics, what characteristics of a document collection make it more amenable to to topic modelling, and how can the potential of topic modellling be harnessed for human consumption (Newman et al. [2010]). The topic coherence

[1] http://mallet.cs.umass.edu/

is measured as

$$score(\omega_i, \omega_j) = log\frac{D(\omega_i, \omega_j) + 1}{D(\omega_i} \tag{4.1}$$

Table 4.1: Top 12 topics

Original	Filtered
program healthy kate lifestyle sessions eat meet programme included organised held healthier encouraging learnt foods beneficial learning relationship handle	program kate lifestyle meet included learnt learning relationship
gp local recently records government copy prescription paper multiple tasmania gps referring avail beginning calls surprised cairns super shared	gp local records government prescription paper gps cairns
physio gp mri injury follow shoulder xray week asked hospital discussed full physiotherapist neck stand complaining neurologist princess forte	physio gp mri xray hospital full physiotherapist neck
call waiting phone told back called list unit rang ring explain apparently assumed clerk calling noticed mcewin lyell requested	call waiting phone back list unit clerk calling
father bed family care appears staff time attending difficult dad speak comfort unit incident law visitor awake gosford palliative	bed family care staff time attending unit palliative
time advised team contact causing consultant tumour professional manner independent safe stressful arrival note closed usual considerate empathetic seizures	time team contact consultant professional manner independent empathetic
looked experience er bad partner full approach worry give free skills male chronic terrible running provider building drive welcomed	looked experience er partner full approach skills male building
night stay thing major hospital admission support suggested accommodation fully sydney unable relatives sleep developed tuesday staff added environment	night stay hospital admission accommodation relatives sleep staff environment
waiting wait hours room hour area waited reception number temperature hurt remember panadol minutes geelong sunday impressed time er	waiting wait hours room area reception number time er
child issues jean aboriginal helped understanding knowledge school minds behaviour hay mighty woods anger louise clinician strategies interaction love	child jean aboriginal understanding knowledge school minds behaviour woods strategies
public brisbane system live mater hospital pa run booking advise toowoomba meds qld health expect west lift weekend weak	public brisbane system hospital run booking meds health west lift
health community sarah local support primary medicare rural group libby people murrumbidgee provide part clients topics art groups guest	health community sarah local primary medicare rural group people murrumbidgee clients guest

Table 4.2: Sum of Composition

Doc Index	Original Composition	Filtered Composition
1	3.920362	3.382284
2	3.742091	3.275150
3	4.396231	3.930039
4	4.275353	3.603821
5	4.569670	4.444751
6	3.153133	2.775077
7	3.368214	2.622281
8	4.541646	4.018145
9	5.399709	5.194911
10	4.498450	3.711763
11	4.255079	4.101976
12	6.755018	6.489172

Table 4.3: Sum of Composition

Doc Index	Original Topic Coherence	Filtered Topic Coherence
1	0	-25.093878
2	0	-32.617172
3	0	-27.143555
4	0	-26.677283
5	0	-28.104525
6	0	-31.384130
7	0	-38.548792
8	0	-41.019321
9	0	-31.220371
10	0	-45.495693
11	0	-52.946951
12	0	-57.020165

Chapter 5

Conclusion and Future Work

Appendix A

Testbed Configuration

Bibliography

M. M. Bismark and D. M. Studdert. Governance of quality of care: a qualitative study of health service boards in victoria, australia. *BMJ quality & safety*, pages bmjqs–2013, 2013.

M. M. Bismark, M. J. Spittal, L. C. Gurrin, M. Ward, and D. M. Studdert. Identification of doctors at risk of recurrent complaints: a national study of healthcare complaints in australia. *BMJ quality & safety*, 22(7):532–540, 2013.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

W. Buntine and A. Jakulin. Applying discrete pca in data analysis. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 59–66. AUAI Press, 2004.

J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Leveraging multi-domain prior knowledge in topic models. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2071–2077. AAAI Press, 2013.

A.-S. Garcia-Moya.L and Berlanga-Llavori.R. Retrieving product features and opinions from customer reviews. *Intelligent Systems*, 28(3):19–27, 2013.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

D. Newman, S. Karimi, L. Cavedon, J. Kay, P. Thomas, and A. Trotman. External evaluation of topic models. In *Australasian Document Computing Symposium (ADCS)*, pages 1–8. School of Information Technologies, University of Sydney, 2009.

D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.