Topic Modelling in Information Quality Research: ICIQ 1996 to ICIQ 2016

(Completed Research Paper)

Jing Gao, School of ITMS, University of South Australia

Philip Woodall, Department of Engineering, Cambridge University

Abstract: Even though new data sources and types have emerged, text based data is still considered to be an important foundation of analysis projects. While text documents deliver important information, they are at the same time difficult to analyse due to their unstructured nature. This research assesses the extent to which cloud-based methods and algorithms are useful to extract information from text based documents for the purpose of Topic Modelling. The chosen subject is the entire ICIQ proceedings from 1996 to 2016. By performing topic modelling on these proceedings, the results also provide a new perspective on how data and information quality research has evolved over time.

Keywords: text analytics, cloud, data quality research

1 INTRODUCTION

The amount of available information has been overflowing during the past years. Organizations generate, receive and store masses of data about customers, employees, orders, suppliers, competitors and many other aspects of their business. This situation is not just a current trend, but is considered to be an ongoing development and the data is expected to grow even faster in the future. Experts predict a continued exponential growth of data increasing roughly 60% annually (Marr, 2014).

Besides the enormous size of data, organizations are challenged by the large number of varying data sources. Sensors integrated into smart phones, cars, credit cards or general consumer electronics all provide a new range of interesting and potentially useful data. Much of this data is available from online sources such as social media websites, which can provide companies with valuable insights about their customers, suppliers and products. The challenge with this type of data is that it is often unstructured, and it is predicted to grow 10-50 times faster compared to structured data (J. Dai, Huang, Huang, Liu, & Sun, 2012, p. 2). Despite being more difficult to process than structured data, unstructured text-based data will continually play a major role for analytics (Russom, 2011, p. 7).

Furthermore, the challenge of unstructured textual data is not just limited to industrial organisations, but also affects academia. With the dramatic increase in the number and availability of academic articles, books, presentations, theses etc. it is increasingly difficult for researchers to stay abreast of the latest topics and trends in an area. Automated tools to support researchers in dealing with information overload have therefore become a critical resource. The area of topic modelling emerged to address this problem, where topic models aim to discover and annotate large archives of documents with thematic information (Blei 2012). Topic modelling could support various types of literature review including, for example, a mapping review. According to Grant and Booth (2009), the purpose of a mapping review is to map out and categorise existing literature according to the quantity and quality of literature, and this could be according to the study design and other key features.

Various text analytics methods now exist that can support this type of analysis in a rapid way, and we attempted to determine the advantages and limitations of these systems when performing topic modelling for a series of academic works. A secondary aim of this paper is to present the results of topic modelling for all papers in the ICIQ conference proceedings published between 1996 and 2016.

This research applies the IBM Watson natural language understanding tool / API (Watson NLU) to the task of topic modelling over a set of academic papers in the data/information quality research area and compares this to the results from expert opinion when performing the same task. The results show what the key topics are in the data/information quality area, and these are contrasted with other similar reviews, and the advantages and limitations of applying an expert system in this context are discussed. The comparison to expert opinion was performed for the papers in ICIQ for the years 2014 and 2015. In addition to the research topics, this research also tries to understand which business domains the published studies belong to. The business domain is useful for researchers to discover because the challenges between different domains within the data/information quality area can differ. Hence, it is convenient for researchers to find all work that has covered the same domain where they are applying their research.

2 BACKGROUND

2.1 Text analytic tools and methods

Among different types of big data, a large proportion of unstructured data is text-based data. Methods to analyse text have been around long before Big Data problems emerged. These methods were the obvious response to many paper transactions being replaced by paperless processes. The transformation of unstructured data from various sources to a highly structured form, which is expected from data mining applications, is the nature of most text analysis processes and a similarity to Big Data projects. There are various open source and commercial tools available for performing text analysis (SAS Text Analytics, IBM SPSS Text Analytics, OpenText, Rapid Miner and others) and many of these have a different focus or specific purpose; these are summarised in Table 1.

Aim/Purpose	Description
Extract Keywords	• Extracts the most important words or word groups from a text to characterize it.
Document Summarization	• Summarizes documents by identifying the most important facts. • Similar to keyword extraction, but is focused on extracting more complex structures (phrases, sentences, paragraphs).
Categorize Documents	• Assigns document to one or more existing categories. • Categories are known to the method.

Cluster Documents	• Automatic organization of documents into groups with similar properties. • Categories and groups will be determined during the analysis process.
Fact Extraction	Identifies facts from digital assets. Answers the question "What happened?"
Concept Extraction (Topic Modelling)	• Clusters groups of words to an idea / concept. • Concepts can be expressed by different words or word groups. • Concepts are not necessarily explicitly stated in the text itself.
Event Extraction	• Identifies events from digital assets. • Answers the question "How did it happen?"
Sentiment Analysis	Detects the attitude of a text with respect to a specific topic or in general. It detects the subjective nature of the text content along with the used tonality (i.e. positive, neutral, and negative).
Extract Entities	• Extracts concrete entities from a text. • E.g. Person -> Usain Bolt, Company -> IBM
Extract Correlations	Facilitates the building of relationships between entities.
Ontology / Taxonomy Modelling	Builds an ontology / taxonomy for a text or a collection of texts.
Making Predictions	Making predictions based on the text content.

Table 1: Functionality of currently available Text Analysis Tools

A topic model represents the abstract "topics" that occur in a collection of documents. Topic modelling is a frequently used text-analytics tool for discovery of hidden semantic structures in a text body. Topic modelling requires building a statistical model to uncover the underlying semantic structure of given documents or text. There are a number of algorithms such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA) can be used for modelling semantics of words based on topics (Hofmann, 2001; Magerman et al, 2001; and Büschken and Allenby, 2016). For example, Griffiths and Steyvers (2004) suggest that LDA can help establish a generative model for documents, which postulates complex latent structures responsible for a set of observations, making it possible to use statistical inference to recover this structure. This approach is particularly useful with text, where the observed data (the words) are explicitly intended to communicate a latent structure (their meaning).

The traditional"manual" topic modelling method often refers to the well-known content analysis method. Berg (2001) defines content analysis as an objective coding scheme applied to data in order to make it amendable to analysis and systematically comparable. According to Busch et al. (1994 - 2012) it is a

"research tool used to determine the presence of certain words or concepts within texts or sets of texts. Researchers quantify and analyse the presence, meanings and relationships of such words and concepts, then make inferences about the messages within the texts". Text in this context is defined very broadly, for instance books, chapters of books essays, interviews, speeches, articles and so on. In order to be able to analyse the content of a text, the text has to be broken down into elements. In the content analysis approach, Zhang and Wildemuth (2005) point out that researchers are often required to develop categorises and a coding scheme (specific words / phrases / ideas belonging to each category)—hence, this is a "manual" approach when the coding scheme is developed by humans . Categories and a coding scheme can be derived from three sources: the data, previous related studies, and theories. However, it is also acknowledged by Zheng and Wildemuth (2005) that performing content analysis can be time-consuming (e.g. to employ multiple coders to ensure consistency), especially when dealing with a large volumes of text.

2.2 Existing Reviews and Textual Analysis of Data Quality Research

During the last decade, researchers have invested time and effort and worked closely with businesses in establishing the data and information quality research domain. Among which, there are a number of projects such as (Zhu et al, 2014, Neely and Cook, 2011, Shankaranarayanan and Blake, 2015), studying how data and information quality research evolves over time. Due to resources and time constraint, each of the mentioned studies derive their conclusions from a sample of available publications. The conclusions are useful to shed the light into the trends in DQ and IQ research.

In practice, researchers often manually perform keywords and concept extraction (topic modelling) analysis for literature reviews (see for example: Wang et al, 1995; Eppler, MJ. and Wittig, 2000). In the 14 different literature review methods, Grant and Booth (2009) point out that grasping business domain and key concepts are essential in review methods such as Mapping review, Rapid review, State-of-the art review, Systematic search and review and Umbrella review. With the development of advance Big Data technologies, in particular, the text analytics capabilities provided on the Cloud-environment such as IBM Watson NLU, it is perhaps possible to re-look at the data quality research trend from the Big data analytics perspective.

3 RESEARCH PROBLEM AND DESIGN

In order to determine the advantages and limitations of applying Natural Language Understanding (NLU) tools to the problem of topic modelling, we chose the IBM Watson Natural Language Understanding API (IBM 2017a) . This was applied to the entire ICIQ conference proceedings published between 1996 and 2016. In particular, the topics were extracted as well as the business domains. A second part of the analysis included a comparison of an expert opinion judgement to the results provided by the NLU tool both for the years 2014 and 2015. During the application of the NLU tool to this topic modelling task, the advantages and limitations of applying the NLU tool were recorded.

Overall, the following approach was used to perform the topic modelling and the comparison (see Figure 1):

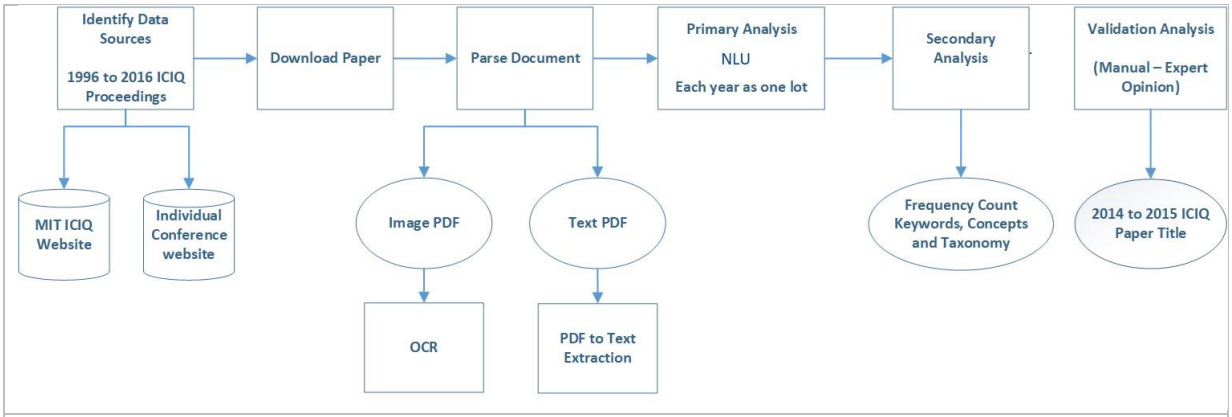

Figure 1: the approach used to perform topic modelling

- Download the ICIQ paper in PDFs from the hosting websites (starting from (http://mitiq.mit.edu/iciq/iqproceedings.aspx);
- Parse the PDF documents to extract text (OCR when required);
- Combine each year's paper into one text file and upload to Watson NLU for processing – Primary analysis; and
- Perform Secondary analysis: Obtain NLU output (per year-base), aggregate extracted keywords (the top 20 were selected based on the relevance score provided by the Watson NLU), concepts and taxonomy from each year and perform a frequency analysis across all years as illustrated in Figure 2 (the same approach was used for the keywords and taxonomy analysis). Although it is possible to combine all years' text together and process in one run (which can be done during the further analysis), the frequency analysis is introduced to assign a equal weighting for each year (assuming each year's conference contributes equally to the overall ICIQ). It is noted that some years' conferences have much more papers than others.

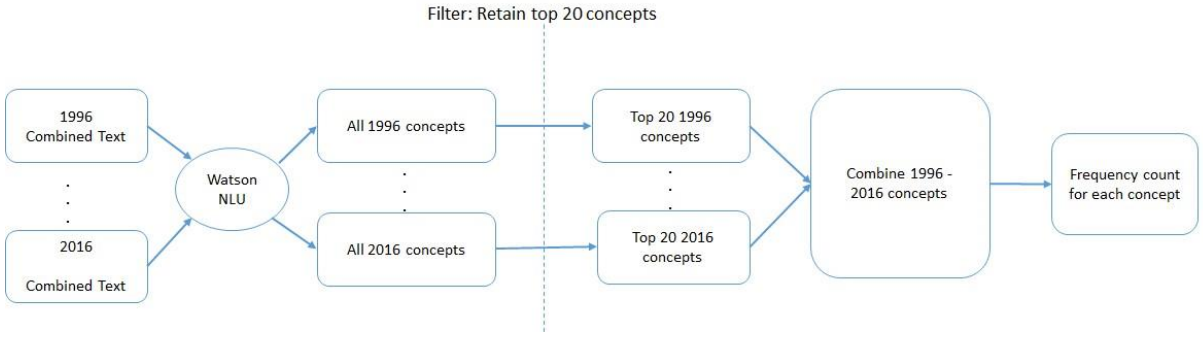

Figure 2: Secondary analysis

- It is noted that during the secondary analysis, the following additional steps were conducted:
    - *Removal of less-informative results:* The results initially contained well-known/lessinformative concepts and keywords (including "Data", "Information", "Quality", "Data Quality". "Information Quality", "Data Quality Management" and "International Conference"). These were removed so that the results only contained the other more useful topics. These were chosen based on two experts (the two authors of this paper) reviewing the results to define a list of less-informative concepts from the

perspective of IQ researchers. This also included removal of less-informative results from the business domain results. Two examples that were removed were "books_and_literature" and "conference".

- *Removal of obvious errors:* the results also contained errors such as topics that were only symbols including "……". This was due to an OCR error. Hence, these needed to be removed.

In addition to the Watson NLU analysis, the researchers decide to take an extra step to conduct a double blinded review process to manually extract concepts (topics) from year 2014 and 2015 paper titles, as the **validation analysis.** Two individual results are discussed after and a consent is reached. The consent is used to compare against the Watson NLU result.

The validation was performed using the following approach:

1. Two researchers (both authors of this paper) independently reviewed the titles of each paper in the ICIQ proceedings (2014 and 2015) and recorded the topic(s) of the paper (note that many papers would contain more than one topic e.g. data mining and cloud computing).
2. For each topic, from the independent reviews, that differed: both researchers compared the answers and came to a consensus agreement over what the actual topic(s) should be.
3. As well as extracting the topics, the above two stages were repeated to extract any relevant business domains that the paper referred to.

Different text analytics may use different wordings for these output terms. For example, IBM Watson NLU produces the above three outputs for a given input (e.g. a text-based document) and refers them as "concepts, "keywords" and "taxonomy". The IBM (2017a) defines the above concepts as:

- Watson NLU **Concept**s: ... high-level concepts that aren't necessarily directly referenced in the text. Concepts are the aggregation of keywords within a given context (e.g. keywords such as data storage and data disposal can be attributed to the concept of data management).
- Watson NLU **Keyword**s: frequently used words or phrases in the text.
- Watson NLU **Taxonomy:** semintics and business domains categorizied by using a five-level classification hierarchy (IBM, 2017b). E.g. /news, /technology_and_computing/software

Note that the term taxonomy is used in the IBM Watson AlchemyLanguage API that this research is based on. On April 7, 2017, the AlchemyLanguage API was upgraded to be known as Natural Language Understanding**.** The term taxonomy has been renamed to category since then.

Depending on interpretations, both keywords and concepts can be considered as topics. For the purpose of the topic modelling in this paper, both the Watson NLU output keywords and concepts were used. Additionally the taxonomy output contains business domain information and so it was used to extract business domains.

4 RESEARCH FINDINGS

This section presents the results of the extraction of topics and business domains from all papers in all years of ICIQ (1996-2016), and then presents the comparison of the results to the expert opinion for papers in ICIQ during 2014 and 2015.

4.1 Top 10 Topics (from concepts) from 1996 to 2016

Key concepts
data_mining
quality_assurance
quality_management
management
data_warehouse
decision_theory
business_intelligence
data_analysis
master_data_management
quality_control

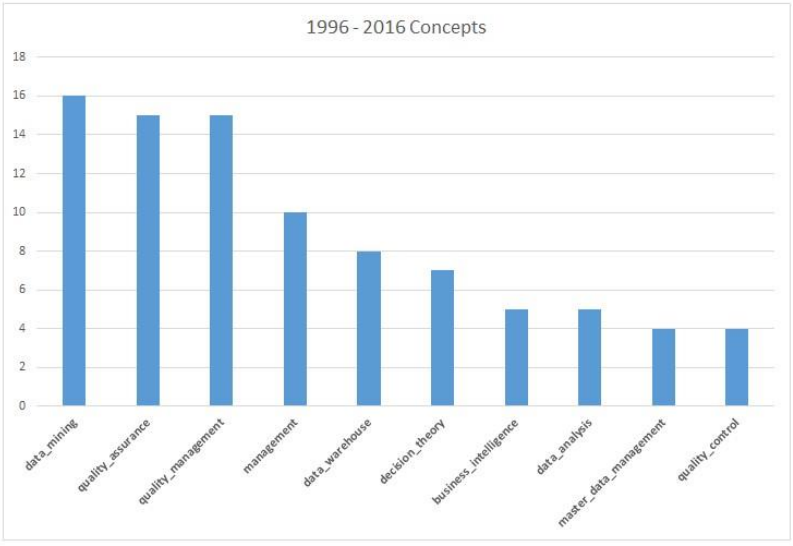

Figure 3: Top 10 topics 1996 –2016, Y scale: Frequency count for each concept, see Figure 2

One of the key problems evident with interpreting the results is that there are no definitions provided by the Watson system. The problem that this raises is that it may be the case that quality assurance is a sub part of quality management or that both of these refer to physical product or service quality management/assurance. In this case it is asserted that quality management/assurance relates to information products etc. rather than physical products. However, we do not know if this is the case or not because no additional data is provided about these answers.

4.2 Top 10 Topics (from keywords) from 1996 to 2016

Keywords
information_systems
data_quality_dimensions
data_quality_assessment
big_data
data_warehouse
information_product
information_integrity_technology
data_elements
information_integrity
social_media_data

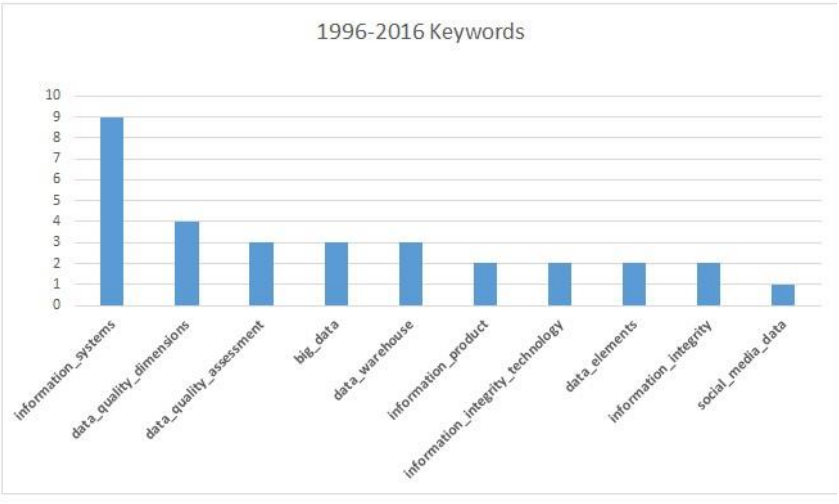

Figure 4: Top 10 ICIQ keywords 1996 – 2016, Y scale: Frequency count for each keyword, refer to Figure 2

When drilling down to the keywords level, it seems that the majority paper has focused on data quality assessment. In addition to the relatively well-established areas such as data warehouse, the new emerging theme is big data including social media data.

4.3 Top 10 Taxonomy categories from 1996 to 2016

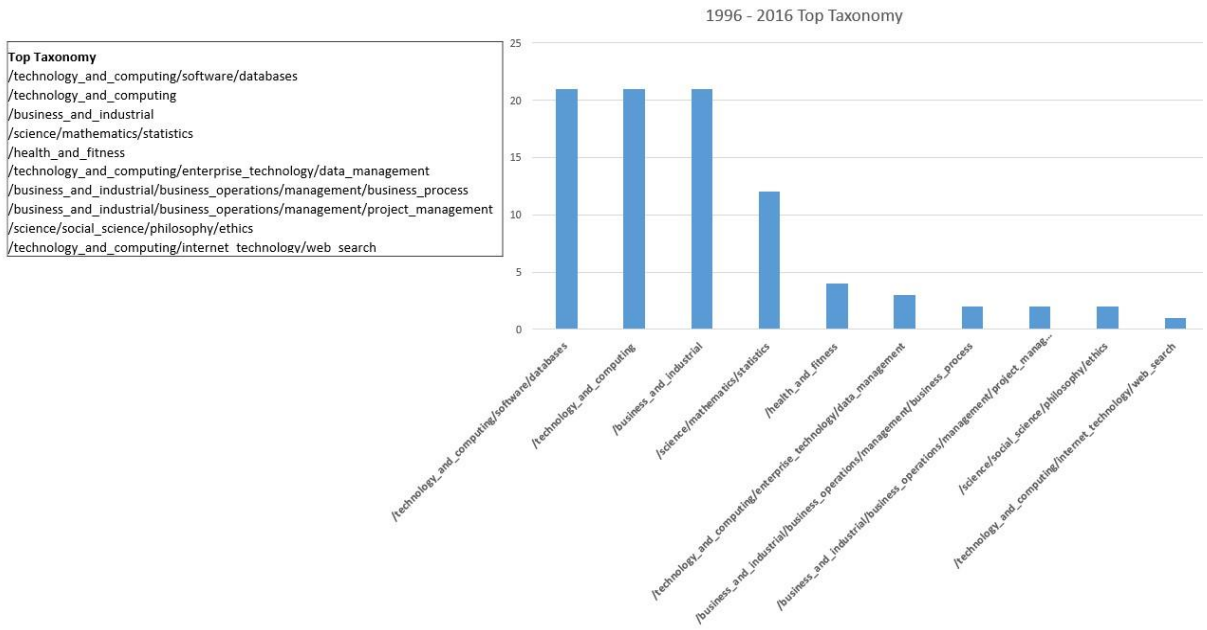

Figure 5: Top 10 ICIQ taxonomy categories 1996 – 2016, Y scale: Frequency count for each taxonomy category, see Figure 2

A clear technology focus on data quality management and assessment can be identified from the above chart. It also shows that the published ICIQ studies are very business drive (business_and_industry appeared multiple times). While the IBM taxonomy categories (IBM 2017b) contains useful results on potential business domains, it is difficult to pin point the actual business domain without an agreed domain classification system. Thus, the identified taxonomy result is presented as it is.

4.4 Validation Analysis - 2014 to 2015

During the validation analysis, the two reviewers in the data quality area extracted topics from the paper titles from the years 2014 and 2015. When conducting the concept extraction, the certain concepts needed to be aggregated and made consistent, therefore the reviewers agreed upon the following coding scheme when combining their results to form a consensus opinion:

- Data analysis includes data analytics, statistical methods;
- Data mining includes machine learning, text mining and topic modelling.
- Entity recognition includes entity resolution and identity resolution.

The same paper titles were processed through the Watson NLU for both the concepts and keywords. Firstly, concepts were analysed, and the following 8 concepts were returned:

- Data mining
- Management

- Data analysis
- Qualitative research
- Data quality
- Business intelligence
- Risk
- Data management

Note that, the Watson NLU also provides a relevance score to rank the results; however, this score was not used in this analysis because we used all of the results that Watson returned, and the focus of the analysis was on coverage/overlaps with the expert opinion rather than priority/rank of the results.

The expert opinion results contained 26 concepts overall, and it must be noted that some concepts were filtered during the analysis. In the case of the expert opinion, certain terms were not considered, and in the case of the Watson NLU results, they were removed. These included, for example, terms such as "Data", "Information", "Quality", "Data Quality". "Information Quality", "Data Quality Management", "International Conference", "Qualitative research", Case study" and "Data Quality Governed". The reason for this is that these terms were considered to be less-informative about the topic areas given that the data source was chosen to be from the data quality/information quality research area.

The second analysis used the keyword results supplied by the Watson NLU, and in this case the potential topic list expanded to 50. A word-to-word direct matching was performed and this yielded four exact matches: Information Quality Projects, Data Mining, Entity Resolution and Data Quality Requirements. However, an inspection of the results indicated that there were other overlaps that were not exact matches but clearly do match semantically. Hence, a semantic match was performed by manually reading each result and making a comparison to the expert opinion results to determine whether they matched or not. The results are shown in Table 2, which includes the list of expert opinion concepts, the concepts from Watson (these results have been grouped by similarly in the table for presentation purposes), and an indication of whether these are considered to be a match or not. Some of the expert opinion results did not match any of the Watson results and these are shown at the bottom of the table. In total, 14 out of the 26 expert opinion concepts did match, 2 were partial matches, and 10 did not match. Partial matches were the cases where the reviewers were not 100% sure that a match exists, however, some of the concepts seemed to overlap. This was the case with "data integration" which seemed to be related to the Watson concepts of Transparent Data Supply, different Data Sources, Information Systems Combination and Multiple Online Sources, which are all related to the challenge of data integration without explicitly stating that the concept is data integration.

Table 2: Watson NLU keywords vs Expert concepts (only in paper titles from 2014 to 2015)

List of topics from the expert opinion	**Match?**	**Matching Watson NLU keywords (grouped by reviewer consent)**
Data Quality Requirements	Yes (direct)	Data Quality Requirements

Data Mining	Yes (direct)	Data Mining Data Mining Application Data Mining Approach Data Mining Method Data Mining Techniques Multi-tiered Medical Data-Mining
Information Quality Projects	Yes (direct)	Information Quality Projects
Entity recognition	Yes (direct)	entity identity information Entity Resolution Event Identity Information identity information management Managing Entity Identity
Data analysis	Yes	Big Data Analytics Customer Classification Analysis
Credit Risk	Yes	Corporate Credit rating Credit Card Credit Card Overdue Credit Card Risk Credit Risk Analysis
Data Cleansing	Yes	Data Cleaning Switch
E-commerce	Yes	E-Commerce Market Based Forecasting E-Commerce Trend
Search Engine	Yes	Generation Search Engine
Information Product	Yes	Information Product Maps True Product Quality
Image Processing	Yes	Hyperspectral Image Denoising
DQ Dimension	Yes	Multi-Dimensional Information Quality
Social Media	Yes	Online Social Communities

Big Data	Yes	Big Data Analytics
Data Integration	Partial	Transparent data supply Different Data Sources Information Systems Combination Multiple Online Sources
Customer Relation Management (CRM)	Partial	Customer Classification Analysis customer data quality
business process improvement	No	
Chief Data officer	No	
Cloud	No	
Crowd sourcing	No	
DQ improvement	No	
DQ Standards	No	
Master data management	No	
Media data	No	
Societal perspective	No	
Visualisation	No	

Two results from Watson appeared to contain an additional word that did not make sense:"Switch" in the Data Cleansing concept and "Generation" in the Search Engine concept. These cases were deemed to be a match without considering the additional strange word.

5 DISCUSSION

Having presented the results, a number of advantages and disadvantages of using Cloud-based topic modelling methods are summarised.

Advantages:

- The ability to process a large volume of data within a short period of time;
- The results are relevant and reasonably accurate, but some manual filtering is still required (e.g. removal of less informative concepts);

- The extraction process is consistent every single time (no human intervention was required, thus avoiding bias); and
- The knowledge-base of cloud-based engine (at least in Watson) is very comprehensive thus meaningful keywords (especially key phrases) can be extracted accurately.

Limitations:

- There are various outputs from these tools (concepts, keywords, and taxonomy, in the Watson NLU case) and it is difficult to determine which of these to use. For example, we found that "ETL" was part of the keywords results, but it was never included in the concepts results. This problem is compounded by the vague documentation on the exact definitions of the output terms. This could present a problem to any researcher wanting to be rigorous in analysing the results.
- The black-box nature of these tools also causes a problem for researchers because it is impossible to determine exactly how the tool has produced the answer. Therefore any assessment of bias or understanding of why certain terms have not been included in the results is very difficult. This is a problem that is likely to persist, as the competitive edge of the tool for the vendors is wrapped up in the algorithms used. Any vendor is therefore very unlikely to want to publish the details of the underlying algorithms.
- Post-processing of the results was required: unlike a manual approach where researchers can identity focus areas (and ignore the rest), the NLU tool processes all text. In our experience, the NLU tool analysed the symbols in the documents (including OCR errors) as well as the words and produced some concepts that were just meaningless symbols. These were clearly no use for the purpose and needed to be removed in order to present sensible results.
- Furthermore, post-processing was also needed to remove obvious results that did not help to understand the content of the ICIQ papers. These included the terms: data quality, data quality management, information quality…
- During different stages of lexical analysis different algorithms can be used (for example to enhance performance for the specific situation/context). However, when using NLU tools it is not possible to switch these algorithms nor understand which algorithms it is using.
- The result may not be repeatable: Cloud-engines such as the Watson NLU use machine learning to constantly fine-tune the processing capability, thus the engine itself evolves over time. Given the same input, there is no guaranty that same results can be produced. (Some may argue that better results may be produced.)

6 CONCLUSION

This research applies the IBM Watson NLU to the task of topic modelling over a set of academic papers in the data/information quality research area and compares this to the results from expert opinion when performing the same task. This Big data topic modelling capability allowed researchers to process the entire collection of ICIQ proceedings from 1996 to 2016. The results show what the key topics are in the data/information quality area, and these are contrasted with other similar reviews, and the advantages and limitations of applying an expert system in this context are discussed. With the Watson NLU results, further analysis can be conducted (e.g. comparing the results against other studies which covers publications in the similar years).

References

Berg, B. L. (2001). QQUALITATIVE RESEARCH METHODS FOR THE SOCIAL SCIENCES (4th Edition). Needham Heights, MA 02494: Allyn & Bacon.

Blei, David, 2012, "Probabilistic Topic Models". Communications of the ACM. 55 (4): 77–84

Busch, C., De Maret, P. S., Flynn, T., Kellum , R., Le, S., Meyers, B., Palmquist, M. (1994 - 2012). Content Analysis. Writing@CSU.

Büschken, J. and Allenby, GM, 2016, Sentence-Based Text Analysis for Customer Reviews, Marketing Science, 2016
Dai, J., Huang, J., Huang, S., Liu, A., & Sun, Y. (2012). THE HADOOP STACK: NEW PARADIGM FOR THE BIG DATA STORAGE AND PROCESSING. Intel Technology Journal, 16(4), 20.
Dickie, H. (1952). ABC Inventory Analysis Shoots for Dollars, not Pennies. Factory Management and Maintenance, 6(109), 92–94.
Eppler, MJ. and Wittig, D., 2000, Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years, Proceedings of the 2000 Conference on Information Quality
Gartner. (2013a). Big Data. Retrieved 22.06.2013, from http://www.gartner.com/it-glossary/big-data/
Grant, M. and Booth, A., 2009, A typology of reviews: An analysis of 14 review types and associated methologies, Health Information & Libraries Journal 26(2):91-108
Griffiths, TL. and Steyvers, M., 2004, Finding scientific topics, Proc Natl Acad Sci U S A. 2004 Apr 6; 101(Suppl 1): 5228–5235.
Gualtieri , M. (2012). THE PRAGMATIC DEFINITION OF BIG DATA. Retrieved 11.08.2013, from http://blogs.forrester.com/mike_gualtieri/12-12-16-technopolitics_podcast_the_pragmatic_definition_of_big_data_explained
Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42(1), pp.177-196.
Howie, T. (2013). The Big Bang: How the Big Data Explosion Is Changing the World. Retrieved 28.07.2013, from http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/04/15/the-bigbang-how-the-big-data-explosion-is-changing-the-world.aspx
IBM, 2017a, Natural Language Understanding Documentation, online accessed [May 15, 2017], URL: https://www.ibm.com/watson/developercloud/doc/natural-language-understanding/
IBM, 2017b, Natural Language Understanding Categories, online accessed [May 15, 2017], URL: https://www.ibm.com/watson/developercloud/doc/natural-language-understanding/categories.html
Kaur, A. and Chopra D., 2016, Comparison of Text Mining Tools, the 5th, International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Sep. 7-9, 2016
Khalilijafarabad, A., Helfert, M. and Ge, M., 2016, Developing a Data Quality Research Taxonomy – an organizational perspective, the proceedings of ICIQ 2016, Ciudad Real, Spain
Koch, R. (1998). Das 80-20-Prinzip. Mehr Erfolg mit weniger Aufwand. Frankfurt am Main: CampusVerlag.
Laney, D. (2001). 3D Data Management: Cotrolling Data Volume, Velocity, and Variety. 4.
Lysons, K., & Farrington, B. (2005). Purchasing and Supply Chain Management: Financal Times Management.
Magerman, T., Van Looy, B., Baesens, B. and Debackere, K., 2011. Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents.
Marr, B. 2014, Big Data: 20 Mind-Boggling Facts Everyone Must Read, Forbes, online accessed [May 12, 2017], URL, https://www.forbes.com/
Neely, P. and Cook, JS., 2011, Fifteen years of data and information quality literature: Developing a research agenda for accounting, Journal of Information systems, Vol 25, No. 1, pp79-108
Russom, P. (2011). Big data analytics. TDWI best practices report, 4, 38
Shankaranarayanan, G. and Blake, R., 2015, Data and Information Quality: Research Themes and Evolving Trends", The 21st American conference on information systems, Puerto Rico, USA
Wang, RY. and Storey, VC. and Firth, CP., 1995, A Framework for analysis of data quality research, IEEE Transactions on knowledge and data engineering, vol 7, no 4, August
Zhang. Y. and Wildemuth, M, 2005, Qualitative Analysis of Content, Analysis 1 (2):1-12
Zhu, H., Madnick, S.E., Lee, Y.W., Wang, R.Y. (2014) Data and Information Quality Research: Its

Evolution and Future, Computing Handbook: Information Systems and Information Technology, 3rd Edition, Editors: Heikki Topi, Allen Tucker. Chapman & Hall / CRC, pp. 16.1-16.20. MITCDO-WP-01