Ore Geology Reviews 111 (2019) 102919

Contents lists available at ScienceDirect

Ore Geology Reviews

journal homepage: www.elsevier.com/locate/oregeorev

GeoDocA – Fast analysis of geological content in mineral exploration reports: A text mining approach☆

Eun-Jung Holden[a,*], Wei Liu[b], Tom Horrocks[a], Rui Wang[b], Daniel Wedge[a], Paul Duuring[c], Trevor Beardsmore[c]

[a] School of Earth Sciences, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia
[b] Dept. of Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia
[c] Geological Survey and Resource Strategy Division, Department of Mines, Industry Regulation and Safety, 100 Plain Street, East Perth, WA 6006, Australia

ARTICLE INFO

Keywords:
Automated document analysis
Geological text mining
Mineral exploration reports

ABSTRACT

Records of past exploration in open-file mineral exploration reports are an important source of information for mineral explorers. These reports document existing geological knowledge that may be relevant to modelling ore forming processes in a particular area of interest. This paper presents the development of GeoDocA, a geological document analysis system, that applies automated text analysis techniques with the specific aim of assisting geologists in browsing of and searching for documents based on relevant geological contents within a large repository of documents. GeoDocA analysed 25,419 exploration reports and using a customised set of keywords pertaining to broad categories such as mineral occurrences, rock types, alteration types, and geological time. An interactive user interface was developed to facilitate visual analysis of exploration reports. For individual reports, it provides a summary of their content in graph form, a gallery of extracted figures and tables, and a list of similar reports based on shared geological keywords. In addition, it assists document search efforts through auto-generated keyword suggestions which are based on associations of keywords learnt by the system from all reports in the repository. While the text mining methods reported here is the foundation for further development to incorporate semantic analysis towards geological knowledge extraction, the outcomes of this study demonstrate the effectiveness of automated text analysis in supporting a fast analysis of a large number of reports to identify the targeted mineral systems and their associated geological environments.

1. Introduction

Publicly available exploration reports are an important source of information for mineral explorers researching geology and mineralisation in their chosen target area. The content in these reports, such as geological ages, existing rock types and types of alteration present, as well as other geological, geochemical and geophysical observations provide key information in mapping and modelling ore forming processes and determining spatial proxies of mineralised systems. However, when a large number of legacy reports exist for a particular mine/area, there is often significant overlap in their content thus it is useful to identify a subset of reports that contain diverse, detailed or anomalous information for further investigation. Selecting reports of interest from a database can be achieved through a first-pass reading of reports or seeking reports containing a specific set of keywords using a document search platform. First-pass reading of multiple reports to analyse complex and diverse details in geology and mining history is time-consuming and subjective. Further, publicly available generic text search tools such as the Google search engine use generalised associations of English words/phrases that are learnt from the web using automated text/knowledge mining technologies. These associations are then used to make auto-completions in the document search, for example, entering `VMS deposit` in the search will suggest a completion word such as `PDF`. For mineral explorers, useful suggestions may be geology/mineralisation specific, e.g. `VMS deposit` and associated rock types typically hosting the type of deposit, geological time scale of the deposition and/or rock alteration types.

This paper reports a geo-specific text mining application, named

☆ Eun-Jung Holden, Tom Horrocks and Daniel Wedge designed and implemented the application workflow and interface. Wei Liu and Rui Wang implemented the natural language processing paradigm and integrated the algorithms. Paul Duuring and Trevor Beardsmore contributed to the design of the application and geological validation of the experimental results.

* Corresponding author.
E-mail address: eun-jung.holden@uwa.edu.au (E.-J. Holden).

https://doi.org/10.1016/j.oregeorev.2019.05.005
Received 21 February 2019; Received in revised form 30 April 2019; Accepted 3 May 2019
Available online 29 May 2019
0169-1368/

GeoDocA, that applied machine analysis to facilitate first-pass reading and searching of reports from a corpus of 25,419 exploration reports using natural language processing, automated text/knowledge mining and visualisation techniques.

1.1. Advances in automated text analysis

In recent years, there have been significant advances in the field of *Natural Language Processing (NLP)*, in particular, knowledge discovery in diverse domains from natural language textual data, often referred to as *text mining* (Peters et al., 2017; Wong et al., 2012; Zhang et al., 2015). The fast paced development in this area even attracted dedicated web sites and data repositories[1] to keep track of its progress.

NLP aims to automatically process natural languages by analysing grammar and sentence structure, with an ultimate goal of understanding semantics in languages, which is the core of human communication. Some downstream or fundamental tasks in NLP include marking up words in sentences as nouns, verbs, adjectives etc., namely part-of-speech (POS) tagging; and segmenting words for Asian languages such as Chinese and Japanese. More advanced and complicated tasks are targeting the semantics in languages such as recognizing predefined categories or entities (Nadeau and Sekine, 2007) such as the names of academics, universities, disciplines etc (i.e. named entity recognition); making sense of relations between entities (Gupta et al., 2017; Yang et al., 2018), for example an academic X working in the university K (i.e. relation extraction); and machine translation and language generation (Sutskever et al., 2014).

Text mining focuses on seeking patterns/structures/trends in unstructured natural language text using both NLP techniques and analytical techniques such as statistical analysis and data mining. Text mining applications vary from automated classification of email messages (Harisinghaney et al., 2014), medical document analysis (Liu et al., 2015) to extraction of concepts in documents (Wang et al., 2015). More broadly, text mining along with structured and semi-structured data mining is an analytical step in discovering knowledge from data.

Knowledge discovery processes typically require end-user interactions to control and interpret the numerical outputs from the analytical steps to extract knowledge. Knowledge Discovery and Data Mining (KDD) is an active research area which encompasses a wide range of analytical techniques (machine learning, artificial intelligence, expert systems etc.) to process structured and unstructured data, with their outputs then visualised and transformed to high level knowledge by accommodating end-user input (Piateski and Frawley, 1991; Frawley et al., 1992).

1.2. Previous research on geological text mining

There have been a number of studies reported in the geological application of text mining. Using a well-known knowledge mining platform called DeepDive (De Sa et al., 2016), Peters et al. (2014) developed PaleoDeepDive which is a machine learning system that detects and extracts paleontological fossil occurrence data from text, tables and figures. The data produced by their machine reading system, which included histories of taxonomic diversity and rates of genus level turnover, was shown to match closely with human generated data. One of the main contributions of this work was using the machine analysis to build a structured database, with an uncertainty estimated for each generated datum. More recently, Peters et al. (2017) reported a combined use of stratigraphic databases, published literature, and a machine reading system to understand the prevalence, extinction, and resurgence of stromatolites in North American marine environments over the past 3 billion years. Their study was able to identify three occurrence phases of stromatolites in geological timescales, as well as predictors for the prevalence of stromatolites—for example, a correlation was found between stromatolite presence and dolomite-bearing marine rock units. Another text mining study on paleontological information was reported by Wang et al. (2018) who built an ontology of geological time scale in North America and integrated fossil occurrences. The extracted information is then visualised geo-spatially.

Very recently, there has been notable interest in Chinese text mining applications for geology. A study by Qiu et al. (2018b,a) reported a Chinese word segmenter that separates semantically and syntactically meaningful words from the geoscience reports. From Chinese text that is a continuous sequence of characters without explicit delimiters, it segmented meaningful words using a deep learning technique, specifically a bi-directional long short-term memory model. Wang et al. (2018) also reported a Chinese word segmentation technique based on a rule-based approach, namely the conditional random fields model that learned the segmentation rules using a customised corpus combining generic and geology terms. These rules were then used to extract geologically relevant terms from reports and a bi-gram association graph of content words were generated based on their co-occurrence frequency. More relevant to the study reported here was the work by Shi et al. (2018) which applied text mining to retrieve mineral deposit related knowledge for the Lala copper deposit in Sichuan Province, China. Convolutional Neural Networks (CNNs) were used to learn and classify four categories of geoscience text data, namely geology, geophysics, geochemistry and remote sensing from Chinese text data. Visualisation of their co-occurrences and frequency statistics were applied to retrieve the associations between the deposit and distinct geological, geophysical, geochemical and remote sensing features. While the above mentioned studies mainly relied on co-occurrence relationships between geological keywords, Enkhsaikhan et al. (2018) reported perhaps the first geological application of semantic analysis. This study carried out word embedding learning to extract semantic similarities between geological phrases as well as semantic relationships between them through an analogy solver.

1.3. The GeoDocA system

This paper presents an on-going development of the Geological Document Analysis system, GeoDocA. Similarly to previous geo-specific applications (Wang et al., 2018; Shi et al., 2018), the system extracts geologically relevant keywords and keyphrases and their co-occurrences in text, and uses visualisation of co-occurrence graphs to retrieve information. However, GeoDocA differs from previous studies in their application with the following specific aims:

(1) to visualise the geological contents within individual reports using: *a summary graph* which represents the occurrences and associations of geological keywords based on co-occurrences in sentences within the report; and *thumbnails of figures and tables* which are automatically extracted from the report;
(2) to identify *a list of similar reports* based on their geological contents using shared geological keywords which were extracted from the reports; and
(3) to provide *search suggestions* for the robust searching of reports from the exploration report repository using the associations of geological keywords learned from the entire database.

Note that our co-occurrence graph is similar to the graph which was referred to as a knowledge graph/tree by Wang et al. (2018) and Shi et al. (2018). Strictly speaking, a knowledge graph requires the recognition of relationships (entity relationships) between keywords/keyphrases (nodes) within the graph, thus the term 'co-occurrence graph' is used throughout this paper.

A preliminary experiment is conducted to analyse the total of 25,419 exploration reports from the Geological Survey of Western Australia's (GSWA) database called WAMEX (Western Australian

[1] https://nlpprogress.com/.

Mineral Exploration reporting system) (Riganti et al., 2015). For geological keywords/phrases, four types of geological entity groups are used, specifically *geological timescales*, *mineralogy*, *host rock types* and *alteration types*. The results show promising and practical outcomes towards using machine analysis to provide a fast, objective and robust document search in the geological domain. Note that the methodologies applied here are expandable to accommodate a wider range of geological keywords, and are adaptable to any geological documents. Our ongoing research focuses on extending the text search framework reported here to incorporate semantic similarities of geological keywords and their relations for knowledge discovery.

The remainder of this paper is structured as follows. Section 2 defines the methods used for extracting text, figures, and tables from the raw PDF reports. Section 3 introduces the text analysis techniques that identify, extract, and rank phrases containing geological context. Section 4 demonstrates an interactive exploration of geological contents in individual reports, and machine supported report search. Section 5 reports on GeoDocA's on-going development towards extracting geological knowledge relevant to ore forming processes. Then a summary is provided in Section 6.

2. Database and pre-processing

2.1. Database

The WAMEX database stores nearly 100,000 digitised or fully digital records and their associated data files submitted by mining companies, which become available to the public after 5 years at no cost (Sunset Clause). Reports are archived as "OCR'd" digital PDFs (only since 2007) with associated digital data. Searching is supported via a web portal (`GeoVIEW.WA`), although each result (i.e., report) must be viewed individually: there is no capacity to preview content nor to search for information within reports. The existing query tool enables searching by the report's *title*, *authors*, *type* (e.g., Annual, Final Surrender), *year of publication*, *target commodity*, *site location* (e.g., unique tenement number), and also by *a list of predefined keywords* (e.g., uranium, amethyst, gravity surveys).

2.2. Pre-processing

The database contains approximately 31,800 exploration reports in PDF format. Each report consists of both textual and graphical content. Since text mining algorithms rely on relatively clean text, graphical content and irrelevant textual information (e.g., references) can inhibit the performance of text mining algorithms. Hence, the goal of pre-processing is to first separate the textual and graphical content of each report, and subsequently remove irrelevant textual information such as invalid characters (e.g., symbols and numbers), tables, bibliographies, references, and tables of contents. Fig. 1 shows a preview of our pre-processing pipeline.

We use the software package *PDFFigs. 2.0* (Clark and Divvala, 2016) to extract and separate textual and graphical content from the reports. For each report, the package returns extracted figures in JPG format, and a JSON file recording hierarchical document information including: the title, section titles and content, title and section positions, and figure and table positions and captions. In mineral exploration reports, these extracted figures often contain key information such as geological maps which illustrate the past or new understanding of geological structures, lithology and stratigraphy of the area, or the source data such as spatial geophysics data from which the geological understanding was derived. Further, high value analytical data such as multi-element geochemistry data observed from rock samples where mineral grades can be identified are often reported in a table. Thus the extraction of the figures and tables is important to understand the content of the report. Figures and tables are extracted using the following three steps:

(1) identifying caption labels using keyword search and predefined rules;
(2) delineating the entire caption based on differentiated text formatting; and
(3) identifying graphical regions and clustering them with the closest caption.

The resulting JSON files are then further processed to remove irrelevant textual content, namely: numbers and symbols; authors, references or bibliographies of reports; figure and table captions; table contents; appendices and attachments; tenement, location, environmental, licensing, and drilling details.

As the reports were written by a variety of exploration companies and there was no requirement to follow a standard template, their document structures and section titles vary. For example, one report may have a section titled `Location and Access`, whereas another may have a similar section named `Location Access`, or `Location and Access Detail`. Since each JSON file recorded all sectional information, we developed a program to parse a JSON file according to section titles. After processing all JSON files, approximately 200 recurring section titles containing irrelevant text were found and recorded on a blacklist. For example, six recurring section titles were found which pertained to location and access and were thus marked as irrelevant, namely: `Location`, `Land Access`, `Tenure and Exploration access`, `Location and Access`, `Location Access`, and `Location and Access Detail`. Numbers or symbols were stripped from section titles by string comparison. Within text, a sentence was excluded if more than 30% of its characters were punctuation marks or numerical characters to remove gibberish sentences; numbers were kept in valid sentences. The final pre-processed text contained section titles and contents for sections with titles not in the blacklist.

3. Identifying, extracting and ranking geological phrases

The pre-processed text is analysed using a standard natural language processing pipeline which consists of:

(1) part-of-speech tagging, which grammatically tags nouns, verbs, and other parts of speech for each word in each sentence;
(2) phrase identification based on tag patterns (e.g., nouns and adjectives, terminated in nouns); and
(3) phrase ranking based on word adjacencies and co-occurrences.In GeoDocA, we extend this process to generate phrases specific to geology by using a dictionary of geological words compiled from Wikipedia and digital geological dictionaries. An example of the geological text mining process is shown in Fig. 2 where a sentence is analysed for parts of speech using Stanford CoreNLP (Manning et al., 2014), which are subsequently used to extract key phrases, which are in-turn filtered by geological content and finally ranked by a phrase ranking algorithm. Each of the text analysis steps outlined above and our method for compiling the geological word dictionary are explained below.

3.1. Part-of-speech tagging

Part-of-speech tagging is a computational process by which parts of speech (e.g., nouns, adjectives, and verbs) are assigned to words in a document or a collection of documents. The CoreNLP package annotates the documents using Penn Treebank POS Tags listed in Table 1. As shown in Fig. 2, after POS tagging, each token in the corpus is annotated with the part-of-speech tags calculated by the tagging algorithms.

3.2. Phrase identification

Syntactic parsing provides us the apparatus to specify patterns that

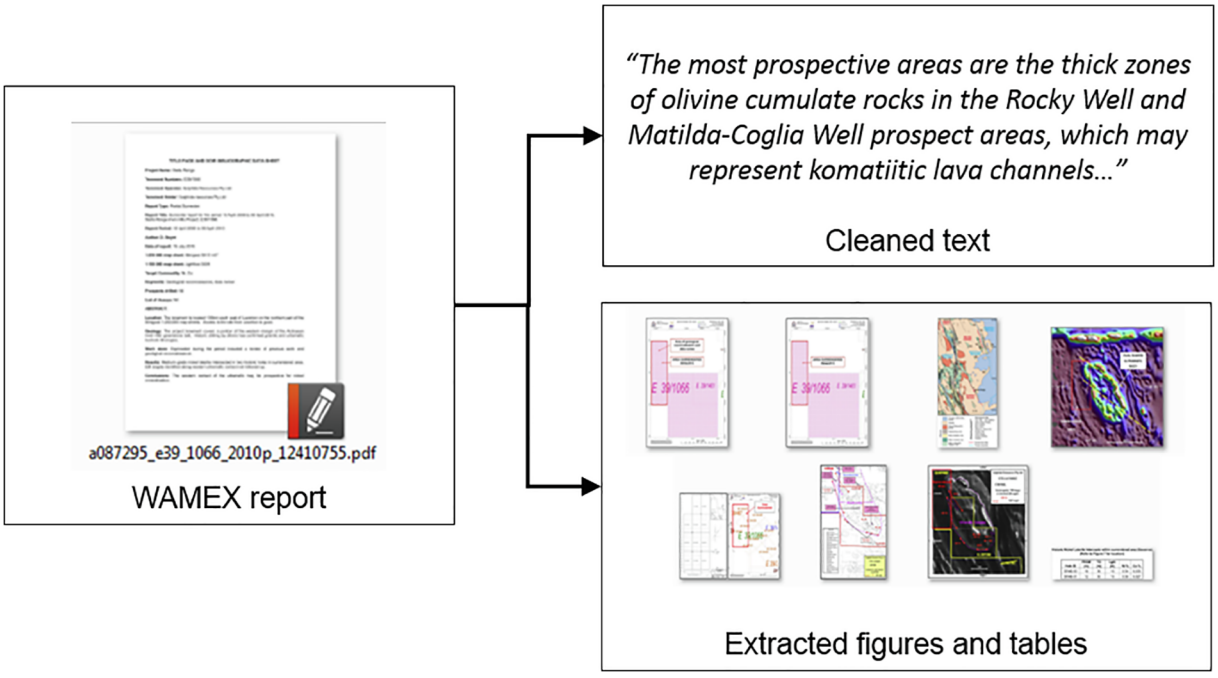


Fig. 1. An example WAMEX report (A087295) split into tables, figures, and cleaned (relevant) text.

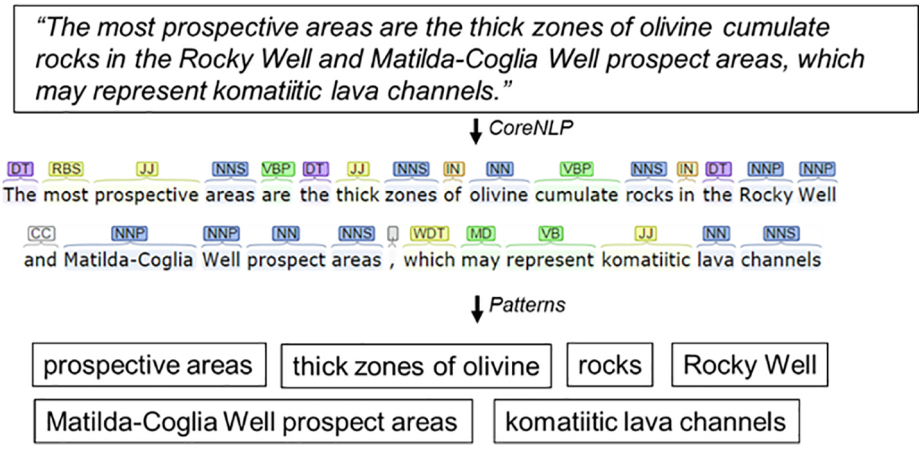


Fig. 2. An example sentence undergoing part-of-speech tagging and phrase extraction.

we can use to perform linguistic filtering. For example, domain concepts or terms typically appear as nouns or noun phrases. Such linguistic filters can be expressed as patterns or rules. The choice of filters can affect how many domain terms extracted. An *open* filter, for example, using the pattern `NP: {<JJ>*<NN.*>+}` permits sequences of nouns following zero or adjectives to be extracted as noun phrases; whereas a *closed* filter such as `NP: {<NN.*>+}` will only extract sequences of nouns.

To further ensure that the extracted candidate terms are in the geological domain, we have constructed a customised dictionary that consists of a list of geological words/phrases. At this point, the dictionary contains words specific to four entity groups which are: geological time scale, mineralogy, rock types and alteration types. The customised dictionary contains the all known minerals (38,705 entries) [2]; geological ages with alternative spellings (159 entries); alteration types (21 entries); and lithology types (from GeMPeT[3], Wikipedia, and the Rocks and Minerals Nature Guide[4] (220 entries)). Note that numerical ages of rocks (e.g. 1530 Ma) are not extracted at this stage of the development.

By consulting the dictionary, the phrases containing a 'geologically relevant' word were kept. The plural form of words is normalised to their corresponding singular form during this process.

3.3. Phrase ranking

Using our geology dictionary an average of 40 keywords were identified per report. These keywords are ranked for visualisation using two algorithms: Term Frequency–Inverse Document Frequency (TF-IDF) (Jones, 1972) and TextRank (Mihalcea and Tarau, 2004a). TF-IDF looks for phrases that are uncommon across all reports, but are frequently used in some reports. TextRank identifies phrases that are commonly used, or are frequently used alongside other commonly used phrases within a single document.

[2] https://www.mindat.org.
[3] http://www.dmp.wa.gov.au/Geoscience-Thesaurus-GeMPet-1564.aspx.
[4] https://www.scribd.com/doc/226163431/Nature-Guide-Rocks-and-Minerals.

Table 1
Penn Treebank POS Tags used in the Stanford Parser (Jurafsky and Martin, 2008)

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	and, but, or	SYM	symbol	+,%, &
CD	cardinal number	one, two	TO	"to"	to
DT	determiner	a, the	UH	interjection	ah, oops
EX	existential 'there'	there	VB	verb base form	eat
FW	foreign word	mea culpa	VBD	verb past tense	ate
IN	preposition/sub-conj	of, in, by	VBG	verb gerund	eating
JJ	adjective	yellow	VBN	verb past participle	eaten
JJR	adj., comparative	bigger	VBP	verb non-3sg pres	eat
JJS	adj., superlative	wildest	VBZ	verb 3sg pres	eats
LS	list item marker	1, 2, One	WDT	wh-determiner	which, that
MD	modal	can, should	WP	wh-pronoun	what, who
NN	noun, sing. or mass	llama	WP$	possessive wh	whose
NNS	noun, plural	llamas	WRB	wh-adverb	how, where
NNP	proper noun, sing.	IBM	$	dollar sign	$
NNPS	proper noun, plural	Carolinas	#	pound sign	#
PDT	predeterminer	all, both	"	left quote	' or "
POS	possessive ending	's	"	right quote	' or "
PRP	personal pronoun	I, you, he	(	left parenthesis	[, (,{, <
PRP$	possessive pronoun	your, one's	)	right parenthesis	+],),}, >
RB	adverb	quickly, never	,	comma	,
RBR	adverb, comparative	faster	.	sentence-final punc	. ! ?
RBS	adverb, superlative	fastest	:	mid-sentence punc	: ; ...– -
RP	particle	up, off			

Keyword	Rating
nickel laterite	1
laterite potential	0.565796
absence of clay development	0.353354
archaean greenstone	0.333367
nickel sulphide	0.327784
stella range area gold values	0.31613
white cliff nickel prospectus	0.31613
sedimentary chert	0.306383
komatiitic lava channels	0.306383
intersection of ni-co laterite intersections	0.306383
hill patch nickel prospect 6km	0.306383
thick zones of olivine	0.306383
greenstone belt	0.305696
amphibolite metamorphic grades	0.287422

Fig. 3. Phrase ranking preview for keywords shown in Fig. 2.

The identified phrases are ranked by both TF-IDF and TextRank, where the final phrases are extracted using the intersection of top ranked phrases delivered by both TF-IDF and TextRank.

3.3.1. TF-IDF

As seen in Fig. 2, a list of candidate phrases are extracted from each document using the linguistic filters. Jones (1972) for the first time defined the concept of term specificity. To determine whether the popularity of a frequent term is due to its significance in the domain or simply because it is a widely used generic phrase across all domains, inverse document frequencies should be used to discount a term's raw frequency. In other words, terms that are only frequent in certain documents are considered more representative with more discriminative power than those that are frequent across the entire document collection.

Let n_i be the number of times term t_i occurred in a document (D), then the term frequency TF_i is calculated as a normalised percentage by the total number of tokens in that document:

$$TF_i = \frac{n_i}{\sum_k n_k}.$$

The inverse document frequency IDF_i is defined based on the ratio of the total number of documents and the number of the documents containing the term:

$$IDF_i = \ln\left(\frac{|D|}{|d: t_i \in d|}\right).$$

The TF-IDF score for term t_i is the product of TF_i and IDF_i. In Fig. 3, these top keywords are displayed with corresponding TF-IDF score shown, normalised by the highest TF-IDF score in the document.

3.3.2. TextRank

TextRank represents a given text as a weighted undirected graph, where each candidate phrase is a vertex and an edge is established if two phrases co-occur within a specified window-size. Given this constructed graph, TextRank provides an authoritative voting mechanism to rank terms according to each node's (i.e., phrase's) degree of connectivity (Mihalcea and Tarau, 2004b). The more connected a phrase is,

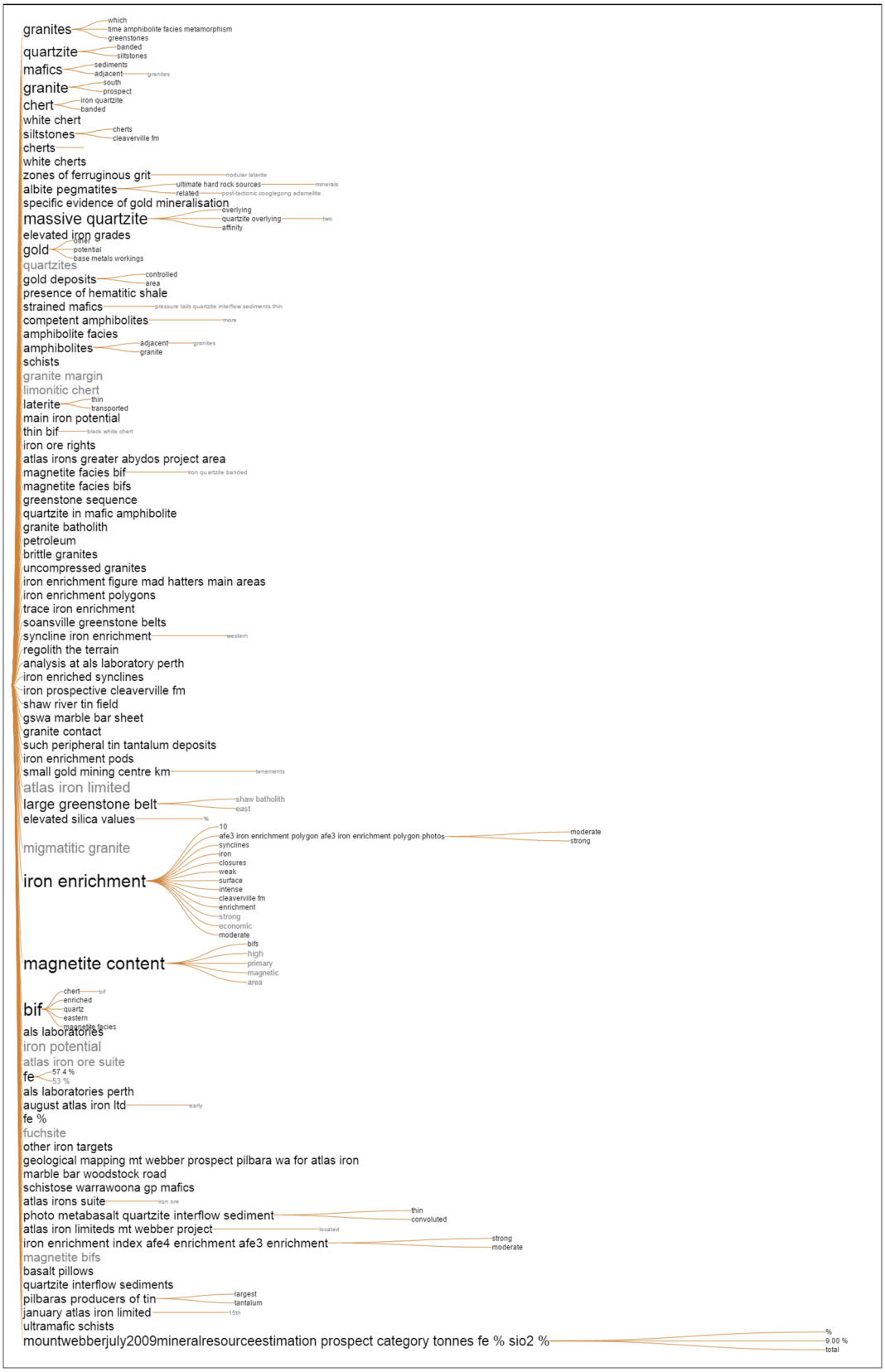


Fig. 4. An example summary graph, displaying identified keywords and their related keywords.

the more important it is.

The ranking score for each vertex is calculated based an adapted version of the PageRank algorithm (Brin and Page, 1998). The original PageRank's voting mechanism is designed for directed unweighted graphs, which calculates the PageRank score of a vertex by its in-degree, i.e., the number of nodes pointing to it (a.k.a. source nodes); the source nodes' scores are further regulated by their out-degrees (the number of nodes to which they are pointing), formally:

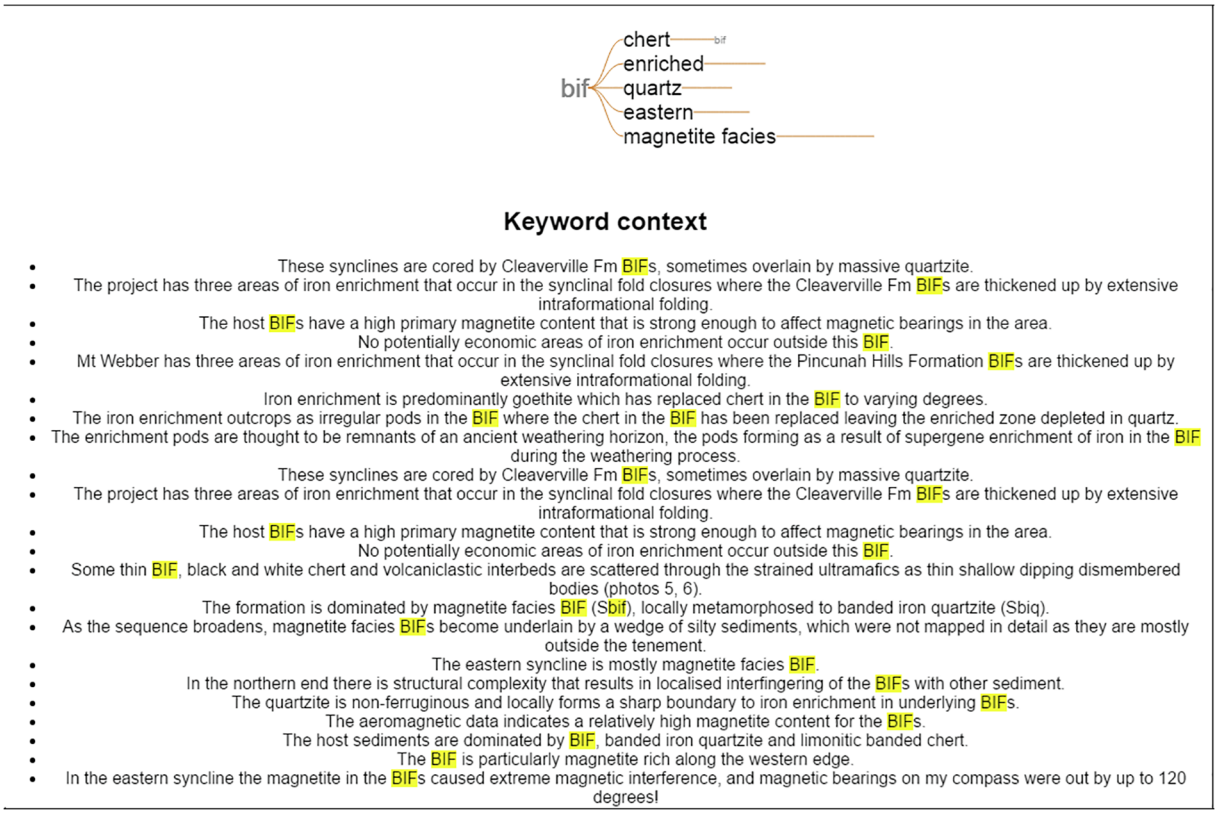


Fig. 5. A node 'bif' selected within the summary graph, with related keywords displayed and extracted occurrences within the report shown with surrounding context.

$$S(v_i) = (1 - d) + d \times \sum_{v_j \in in(v_i)} \frac{1}{|out(v_j)|} S(v_j),$$

where d is damping factor, often set empirically to 0.85 (Mihalcea and Tarau, 2004b), $in(v_i)$ is the set of vertices pointing to vertex v_i, and $out(v_i)$ is the set of vertices to which v_i is pointing.

For an undirected graph, a vertex's in-degree and out-degree are considered the same. The adapted TextRank algorithm is then:

$$WS(v_i) = (1 - d) + d \times \sum_{v_j \in in(v_i)} \frac{w_{ij}}{\sum_{v_k \in in(v_j)} w_{jk}} WS(v_j),$$

where w_{ij} is the strength of connection between two vertices, v_i and v_j.

4. Experimental results on analysis and search of reports

The GeoDocA has an interface that supports interactive visualisation of machine reading outputs and using them to search for documents. Given the geological keyphrases extracted using the techniques described in the previous section, it generates summary graphs and extracts figures and tables from individual reports; identifies similar reports solely based on occurrence of geological keywords within reports; and provides auto-completion suggestions based on keyphrase co-occurrences from the entire corpus.

4.1. A summary graph

A summary graph consists of geological phrases that are extracted from the report, and their associations based on their co-occurrence. This graph is generated by grammatically re-parsing the sentence and looking for words which interact with our geological phrases using the Stanford CoreNLP implementation of 'dependency parse trees' (de Marneffe et al., 2006). Fig. 4 shows the summary graph of an example report, namely a Mt. Webber mineral exploration report. In the summary graph, a node can be interactively chosen to see its association with other keywords. When interactively selecting a node representing 'bif' (banded iron-formation), which is a common keyword in the report, Fig. 5 shows the child terms identified in the summary graph, which are geologically acceptable outputs because of their common genetic, compositional, and/or spatial affiliation with banded iron-formations in a natural geological setting.

4.2. Figures and tables

Figures and tables are extracted from reports by identifying their captions using an existing approach used in Semantic Scholar, namely PDFFigs. 2.0 (Clark and Divvala, 2016). The extracted figures and tables are presented in a list to allow rapid viewing of maps, other figures and tables to rapidly identify visual content. This is useful when searching through a large number of documents for a specific map or set of results.

Fig. 6 shows the analysis output of a report "Mount Webber E45/2268". This example shows that GeoDocA displays the thumbnails of figures and tables that are automatically extracted from the report, and lists the most similar documents based on keywords.

4.3. Similar documents

For a selected document, similar documents are found and ranked as shown in Fig. 6. Firstly, the top 10 keywords (*top-10*) for the selected document (according to TF-IDF or TextRank) are identified, and a document index is used to identify other documents containing any of these keywords. Then for each identified document, the number of keywords in common with the top-10 and the sum of their corresponding scores is recorded. Identified documents are then ranked by the number of top-10 keywords; in the event of a tie, they are ranked by

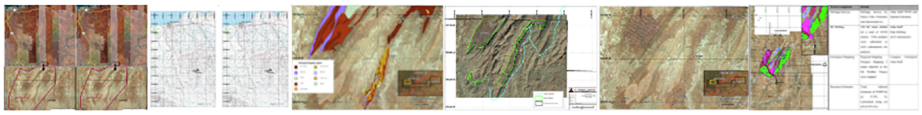
Mount Webber E45/2268, Annual Report for the period 30/01/2009 – 29/01/2010
A-number: 086431

View original PDF

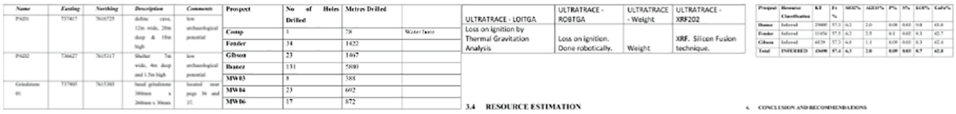
Abstract

Annual Report - Mount Webber (30/01/2009 - 29/01/2010)

List of figures and tables



Tenement	Holder	Date Granted	Area (Blocks)
E45/2268	Attgold Pty Ltd	30/01/2009	18



Most similar documents by keyword

- 108443: Mount Webber Project E45/2312-I FINAL SURRENDER REPORT To the Department of Mines and Petroleum For the period From 29 September 2006 to 11December 2015
- 091621: Mt Webber Project E45/2312 Annual Technical Report to DMP Period Ending 28 September 2011
- 105761: ANNUAL REPORT For the Period 27 March 2014 to 26 March 2015 MT WEBBER PROJECT E45/2288
- 108233: MT WEBBER PROJECT E45/2288 Final Surrender Report To the Department of Mines and Petroleum For the Period From 27 March 2014 to 10December 2015
- 088648: Mt Webber Project, Annual Report to the Department of Mines and Petroleum for the period 29/09/09 to 28/09/10, E45/2312-I
- 107321: MOUNT WEBBER PROJECT Partial Surrender Report to the Department of Mines and Petroleum for the Period 29 September 2009 To 27 August 2015
- 078725: Annual Report to the Department of Industry and Resources Abydos for the period 01/05/07 to 30/05/08, TENEMENTS: E45/2362, E45/2363, E45/2404, E45/2471, E45/2594, E45/2899 & E45/2765, Combined Reporting: C40/2007
- 078646: ANNUAL REPORT to the Department of Industry and Resources for the Period 24/04/07 to 23/04/08 Tenement: E45/2728
- 082658: Annual Report to the Department of Mines and Petroleum for the period 01/05/2008 to 30/04/2009 Tenements: E45/2362, E45/2363, E45/2404, E45/2471, E45/2594, E45/2899 & E45/2765, Combined Reporting: C40/2007.
- 089997: Wodgina Project, Annual Report for the period ending 6th March 2011, P45/2598.

Fig. 6. An example report analysis, with extracted figures that can be viewed and downloaded by clicking on auto-generated hyperlinks. GeoDocA recommends related documents that have similar geological contents.

the total score for those keywords. Only the top 10 keywords are chosen as a trade-off between the quality of the similarity measure and processing time.

The results of the automated analysis of an example document are shown in Fig. 6. Report A-086431 documents exploration work completed by Atlas Iron Ltd. in the Mount Webber project area during 2009. The report also reviews past exploration in the area and presents a summary of general geological information, such as its geographical location and a description of major rock types in the region. A large portion of the document describes the details of exploration completed during the reporting period, including heritage surveys, geological mapping, drilling results, and resource estimations. Exceptionally, this exploration report includes a detailed review of geological mapping in the Mt Webber project area by an external consultant to Atlas Iron Ltd. This report, included in the appendices, provides a well-crafted and detailed description of the local geology, structure, alteration, and iron mineralization in the area, with inclusion of multiple maps and photographs. An additional appendix presents resource estimations for the Mt Webber project area based on drilling information, with minor repetition about the project geology from the previous geological mapping section of the appendices. The digital appendices include text document information for drilling results, high-resolution copies of some of the figures included in the main body of the A-086431 report, as well as relevant files for projecting geological mapping data in the MapInfo GIS program.

The recommended similar reports included in the document list labelled “most similar documents by keyword” are a reasonable query output, firstly based on the inclusion of “Mount Webber” in the titles of the first six returned reports and secondly considering the contents of these reports. The highest recommended report A-108443, considered by our query to be most similar to the targeted report A-086431, is a surrender report by Atlas Iron Ltd. to the Department of Mines and Petroleum in 2016. It presents a summary of all work completed by this company from 2006 to 2015. Consequently, this report represents a good starting point of reference for users of the WAMEX database for querying details about the geology and work completed on the Mount Webber project area. The second highest recommended report A-91621 (submitted in 2011), employs common report subject headings, including an introduction to the regional and local geology of the Mount Webber project area, history of exploration, and current exploration activities. It represents a useful literature recommendation for interested readers because it displays relevant subject matter without significant plagiarism or repetition of data from the previously described target, or highest recommended, reports. The lower ranked reports that contain Mount Webber in their titles (e.g. A-105761, A-108233, A-088648, and A-107321) are all relevant suggestions for the search about information for the project area. They are all reports compiled by Atlas Iron Ltd. and use similar report headings and summarize similar geological information. Their broad similarity in style and content implies that they are likely to be of similar value to a user. The seventh

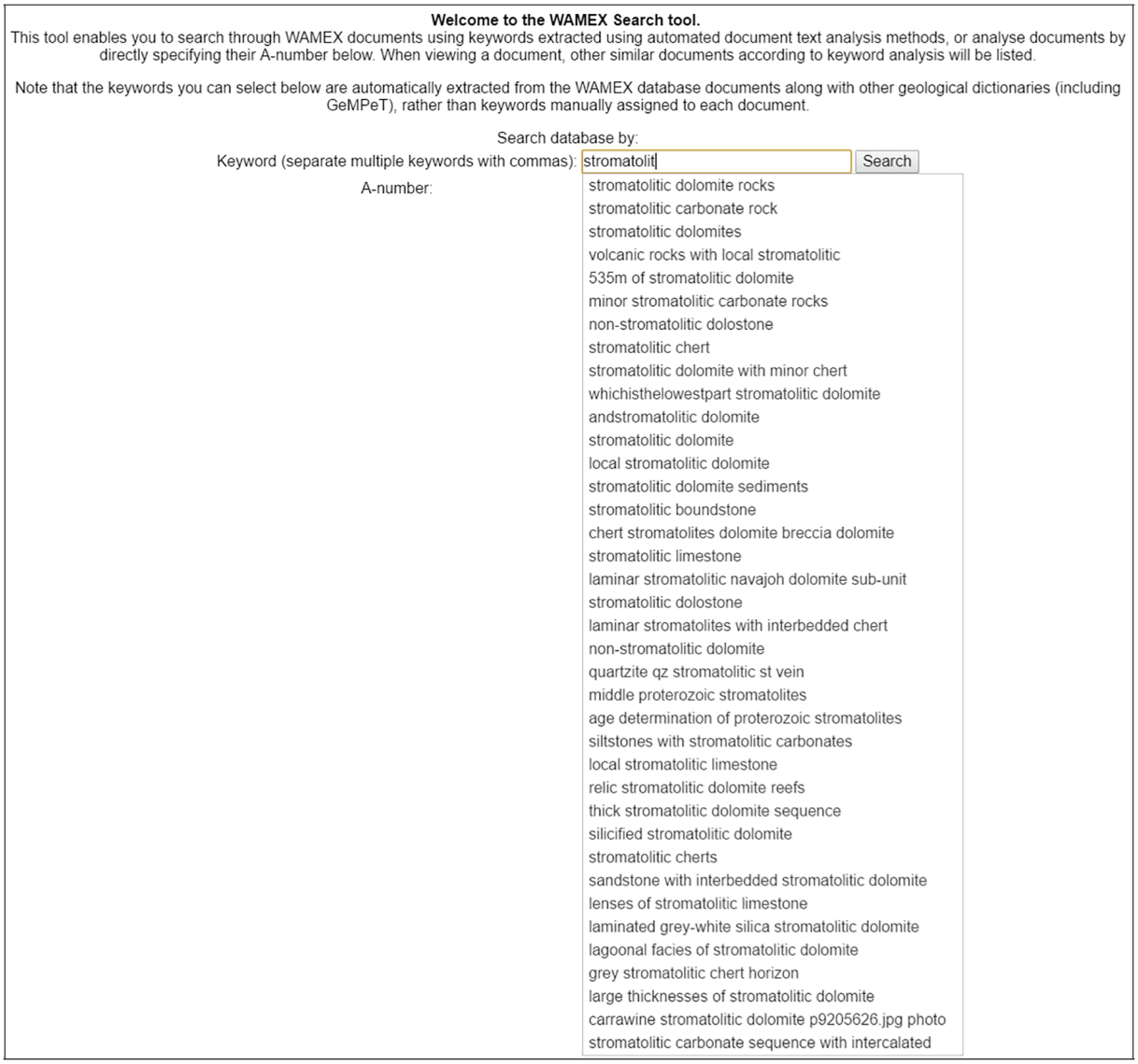


Fig. 7. Auto-suggested completions for keyword search based on the partial entry of 'stromatolit'.

to tenth recommended reports do not share the Mount Webber locational term in the title of these reports. For example, the seventh and tenth recommended reports (A-078725 and A-089997) are summaries of work reported for the distal Abydos and Wodgina project areas, located 48 and 75 km away from Mount Webber, respectively. These reports include reviews of the regional geology in the east Pilbara area, which also includes the Mount Webber project area. Apart from this overlap in content, these low-ranked recommended reports are considered to be only vaguely relevant to Mount Webber.

4.4. Auto-generated keyword suggestions

Reports can be searched through the report number in the database but alternatively a set of geological keywords can be used. In the keyword search, GeoDocA provides auto-completions using the co-occurrences learnt from all of the reports in the respository. Fig. 7 shows a user input of 'stromatolit' (allowing for matches for both 'stromatolite' and 'stromatolitic') and the resulting auto-generated suggestions. Previously Peters et al. (2017) reported in their text mining research using stratigraphic databases about the associations between prevalence of stromatolites and dolomite bearing marine rock units. In the figure, auto-suggestions for 'stromatolites' include 'dolomites' in different combinations of keywords, which shows the co-occurrences of stromatolites and dolomites in reports within the WAMEX database.

5. On-going & future developments

GeoDocA is still under development with various improvements to be made. This prototype uses only a small subset of geological keywords useful for mineral exploration, but this customised set can be expanded, e.g. incorporating different types of survey/analytical data mentioned in reports, or to build a more generic geological document analysis.

The main on-going development is to extend GeoDocA towards extracting geological knowledge about mineralisation, e.g. extracting information about mineral deposit associations, the favorable geological environment in which they form, and the key genetic processes involved in their formation. To achieve this, we will focus on building a knowledge graph, which differs from the summary graph presented in this paper that is based on keyword co-occurrences only. More sophisticated text mining techniques are required to provide meaningful connections between keywords in the knowledge graph. For example, 'rock type X' (node) is 'altered' (connection) to become 'rock type Y' (node) which 'hosts' (connection) 'mineralisation M'. This process needs to incorporate entity linking that automatically links geological

keywords to their entity types. Further, refining a knowledge graph to incorporate new information requires semantic similarity comparison between the new and the existing information in the knowledge graph. We reported progress in this area of research (Enkhsaikhan et al., 2018), which involves automatically annotating raw text (geological keywords) using known entity groups such as 'rock type', 'location', 'commodity' or 'stratigraphic unit'. In addition, widely-used word embedding analysis techniques are applied to establish semantic relationships between geological keywords. As a result, two potential applications were reported. One is searching for semantically associated keywords. For example, given 'ashburton formation', the prototype system generated 'wyloo group', 'mount minnie group', 'mount mcgraph formation', 'june hill volcanics' and 'capricorn group'. The relevance of these extracted keywords was validated by a database that was not used to train the semantic analysis network. From a mineral exploration perspective, such analysis is useful to understand stratigraphic relationships which may be one of the controls on mineral deposition. Another application is an analogy test. Using annotated entity types and semantic relations, an association between two entity groups such as 'location' and 'commodity' can be established and used for Question & Answering (Q&A). An example may be *What is the location of 'iron ore (deposits)', which has the same relations between 'Kalgoorlie' and 'archean gold (deposits)'?*. Such tools can assist mineral explorers towards understanding and establishing relationships between different types of mineral deposits, their locations and geological environments. This level of knowledge mining from geological texts is still at its infancy. Once developed, this Q&A system may be used to automatically extract the knowledge of mineral deposits, their depositional environments, the processes and their spatial distributions, which can potentially be used to validate or design a mineral exploration framework in the future. Further, mineral potential mapping is typically performed in a geo-spatial context (Porwal and Kreuzer, 2010). Thus retrieving the location(s) where the specific geological knowledge is extracted from and visualising them in a GIS environment will be an important aspect for future development.

6. Summary

This paper reports GeoDocA, which was developed to assist the search and analysis of reports based on their geological contents. Using a customised set of domain specific keywords, the system adapts and extends the existing natural language and text mining technologies to provide a practical tool for fast browsing of reports based on their geological content while supporting auto-completion driven search of reports. The summary of contents based on geological keywords and their co-occurrence based associations in an interactive graph form is an effective way to explore the contents of the report. Further, GeoDocA also provides the visualisation of the figures and tables within the selected report, and the identified similar reports in the repository based on geological keywords, offering an effective way to utilise text data to seek geological information. On-going developments of GeoDocA aim to build a knowledge graph from which semantically meaningful associations of the keywords can be extracted. Such extension may assist the automated identification of geological environments and processes favorable for mineralisation to establish a robust mineral exploration framework. As the first step towards this goal, the preliminary study reported here demonstrates that co-occurrence based associations of geological keywords can be effectively used to assist mineral explorers for fast browsing of and robust search for reports from a large repository of exploration reports.

Computer code availability

A video demonstrating GeoDocA is available online at: https://youtu.be/PBbDhRslBNY

This study used the following public domain source codes:

- PDFFigs. 2.0:http://pdffigures2.allenai.org/
- Standford CoreNLP:https://stanfordnlp.github.io/CoreNLP/
- Gensim (Term Frequency-Inverse Document Frequency (TF-IDF)) :https://radimrehurek.com/gensim/
- NetworkX (TextRank):https://networkx.github.io/

Acknowledgement

This research was funded by the Geological Survey of Western Australia through the State Government's Exploration Incentive Scheme. Paul Duuring and Trevor Beardsmore publish with permission from the Executive Director of the Geological Survey of Western Australia.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.oregeorev.2019.05.005.

References

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. Computer Netw. ISDN Syst. 30 (1–7), 107–117.
Clark, C., Divvala, S., 2016. 2.0: Mining figures from research papers. In: IEEE/ACM Joint Conference on Digital Libraries (JCDL) IEEE, pp. 143–152.
de Marneffe, M.-C., MacCartney, B., Manning, C.D., 2006. Generating typed dependency parses from phrase structure parses. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC).
De Sa, C., Ratner, A., Ré, C., Shin, J., Wang, F., Wu, S., Zhang, C., Jun. 2016. Deepdive: declarative knowledge base construction. SIGMOD Rec. 45 (1), 60–67.
Enkhsaikhan, M., Liu, W., Holden, E.-J., Duuring, P., 2018. Towards geological knowledge discovery using vector-based semantic similarity. In: Proceedings of the International Conference on Advanced Data Mining and Applications. Springer, Cham, pp. 224–237.
Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J., 1992. Knowledge discovery in databases: an overview. AI Mag. 13 (3), 57.
Gupta, N., Singh, S., Roth, D., 2017. Entity linking via joint encoding of types, descriptions, and context. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmarkpp. 2681–2690. https://www.aclweb.org/anthology/D17-1284.
Harisinghaney, A., Dixit, A., Gupta, S., Arora, A., 2014. Text and image based spam email classification using KNN, Naïve Bayes and reverse DBSCAN algorithm. In: Proceedings of International Conference on Optimization, Reliabilty, and Information Technology (ICROIT). IEEE, pp. 153–155.
Jones, K.S., 1972. A statistical interpretation of term specificity and its applications in retrieval. J. Doc. 28 (1), 11–21.
Jurafsky, D., Martin, J.H., 2008. Speech and Language Processing (Chapter 10). Pearson Education (US).
Liu, W., Chung, B.C., Wang, R., Ng, J., Morlet, N., 2015. A genetic algorithm enabled ensemble for unsupervised medical term extraction from clinical letters. Health Inform. Sci. Syst. 3 (1), 1–14.
Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D., 2014. The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55–60.
Mihalcea, R., Tarau, P., 2004a. Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing.
Mihalcea, R., Tarau, P., 2004b. Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.
Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Linguisticae Investigationes 30 (1), 3–26 Publisher: John Benjamins Publishing Company.
Peters, S.E., Husson, J.M., Wilcots, J., 2017. The rise and fall of stromatolites in shallow marine environments. Geology 45 (6), 487.
Peters, S.E., Zhang, C., Livny, M., Ré, C., 2014. A machine reading system for assembling synthetic paleontological databases. PLOS ONE 9 (12), 1–22.
Piateski, G., Frawley, W., 1991. Knowledge Discovery in Databases. MIT press.
Porwal, A.K., Kreuzer, O.P., 2010. Introduction to the special issue: mineral prospectivity analysis and quantitative resource estimation. Ore Geol. Rev. 38 (3), 121–127.
Qiu, Q., Xie, Z., Liang, W., 2018a. A cyclic self-learning chinese word segmenation for the geoscience domain. Geomatica 72, 16–26.
Qiu, Q., Xie, Z., Liang, W., Wenjia, L., 2018b. Dgeosegmenter: a dictionary-based chinese word segmenter for the geoscience domain. Comput. Geosci. 121, 1–11.
Riganti, A., Farrell, T.R., Ellis, M.J., Irimies, F., Strickland, C.D., Martin, S.K., Wallace, D.J., 2015. 125years of legacy data at the geological survey of western australia: capture and delivery. GeoResJ 6, 175–194.
Shi, L., Jiangping, C., Jie, X., 2018. Prospecting information extraction by text mining based on convolutional neural networks-a case study of the Lala copper deposit. IEEE

Access, China.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14 MIT Press, Cambridge, MA, USA, pp. 3104–3112. http://dl.acm.org/citation.cfm?id=2969033.2969173.
Wang, C., Ma, X., Chen, J., 2018. Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. Comput. Geosci. 115, 12–19.
Wang, C., Ma, X., Chen, J., Chen, J., 2018. Information extraction and knowledge graph construction from geoscience literature. Comput. Geosci. 112, 112–120.
Wang, R., Liu, W., McDonald, C., 2015. Using word embeddings to enhance keyword identification for scientific publications. In: Databases Theory and Applications. Springer, pp. 257–268.
Wong, W., Liu, W., Bennamoun, M., 2012. Ontology learning from text: a look back and into the future. ACM Comput. Surveys (CSUR) 44 (4), 20.
Yang, D., Wang, S., Li, Z., 2018. Ensemble neural relation extraction with adaptive boosting. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. IJCAI'18 AAAI Press, pp. 4532–4538. http://dl.acm.org/citation.cfm?id=3304222.3304400.
Zhang, Y., Chen, M., Liu, L., 2015. A review on text mining. In: Proceedings of the 6th IEEE International Conference on Software Engineering and Service Science (ICSESS) IEEE, pp. 681–685.