Term Frequency Analysis for Semantic Modeling of Geological Fault Knowledge in the Energy Industry

Fabio C. Cordeiro[1,3,*,†], Yuanwei Qu[2,†]

[1]Petrobras Research and Development Center, Avenida Horácio de Macedo, 950, 21941-915, Rio de Janeiro, Brazil
[2]SIRIUS center, Department of Informatics, University of Oslo, Gaustadalléen 23B, 0373 Oslo, Norway
[3]Getulio Vargas Foundation, Praia de Botafogo, 190, 22250-900, Rio de Janeiro, Brazil

Abstract

Understanding geological faults is crucial for the oil and gas industry, as it affects the production performance of reservoirs. Nevertheless, the fragmented and ambiguous nature of geological fault information hinders efficient information retrieval. Formal geological ontologies offer a solution by enabling domain-specific data integration. One challenge that persists in ontology development is defining a set of relevant terms with good coverage used in the domain community. Based on the TF-IDF method, we conduct a term frequency study of fault-related concepts in recent academic paper abstracts. We select papers from diverse journals and evaluate terms with geologists and ontologists. The results align with experts' knowledge and contribute to the construction of a vocabulary list for the geological fault knowledge model and pave the path for thorough ontological analyses of geological faults, which facilitates data retrieval and mitigates semantic ambiguity. Future work includes improving the quality of the generated vocabulary list, implementing the proposed corpus internally, and considering more in-house technical documents for a more comprehensive coverage.

Keywords

Term Frequency Analysis, Geological Fault, Knowledge Modeling, Ontology Development

1. Introduction

A comprehensive understanding of a geological fault is crucial for the oil and gas industry, as it directly influences reservoir quality, potentially leading to leaks or maintaining a seal [1]. In addition to the oil and gas industry, faults also play an essential role in mining, geothermal, and construction industries [2]. However, the geological data required for interpretation is often scattered across various sources and managed by different disciplines, presenting a complex challenge for geological information retrieval [3]. Furthermore, the geological knowledge derived from such dispersed and sparse data is often fraught with ambiguity [4].

The term 'fault' can represent a spatial arrangement structure, an abstract 2D plane, or a 3D deformed volume [5]. However, in textual documents, all these concepts are expressed simply as 'fault.' This ambiguity in geological fault knowledge and terminology significantly hinders the efficiency of geological information retrieval from complex databases. Consequently, there is a

SemIIM'23: 2nd International Workshop on Semantic Industrial Information Modelling, 7th November 2023, Athens, Greece, co-located with 22nd International Semantic Web Conference (ISWC 2023)

*Corresponding author.
†These authors contributed equally.
fabio.cordeiro@petrobras.com.br (F. C. Cordeiro); quy@ifi.uio.no (Y. Qu)

growing demand within the oil and gas industry for integrated geological data and information models capable of enhancing the retrieval process. To address this demand, one key solution is the formalization of geological knowledge.

Building formal geological ontologies stands out as a promising solution for domain-specific data integration and retrieval in the oil and gas industry. In semantic technologies, an ontology is a formal (machine-readable), explicit specification of a conceptualization (abstract of the world that we want to represent in the knowledge) that is shared and agreed upon by the domain community [6]. These ontologies establish a semantic foundation that enhances search engines' capability to recognize and interpret domain-specific terminology and relationships.

Within the geological community, various ontologies have been proposed, ranging from core-ontology for geology [7], fracture ontology [8], plastic rock deformation ontology [9], geological map ontology [10], geological time ontology [11], fault ontology [5] ,deep-marine deposits [12], to risks associated with the petroleum reservoir [13]. A notable industrial case is Petrobras, the Brazilian oil company, which is developing a specialized search engine for geoscientific technical reports empowered by ontology and knowledge graphs [14, 15].

The development of an ontology requires the involvement of ontologists, domain experts and users to define the purpose, scope, requirements, etc [16]. During the conceptualization and formalization stages, domain experts bear the primary responsibility for selecting and providing knowledge and terminology resources. Yet, assessing the comprehensiveness of the selected terms and concepts, particularly in a domain as semantically ambiguous as geology, poses challenges. Furthermore, there is also a need to convince non-domain users that the selected terms have good coverage of domain knowledge, which is generally accepted within the geology community. To address this challenge during ontology development, approaches such as term frequency analysis and information extraction from domain documents are recommended to employ [17]. This method has been applied in various geological information and knowledge modeling tasks, such as the subsurface energy [18], mineral exploration [19, 20], and geological natural hazard [21]. However, the essential yet ambiguous concept of 'fault' has not yet undergone a term frequency analysis.

In this paper, to support the knowledge modeling of geological fault for the energy industry, we conduct a term frequency study (Sect. 2) of the 'fault' concept and its related terms from academic paper abstracts. Compared to the verbose full paper, the abstract contains the most important concepts of the research objectives. Cao et al. [22] compared topics extracted from academic papers abstracts and full text and found that the similarity between results is higher when more documents are analyzed. To balance the focus and extension of fault concept, the selected papers range from domain-specific, domain-related, and industrial-related to general domain journals. We listed the renowned geoscience journals with good impact factors and citation scores, and then specialists chose the most relevant for each domain. The extracted terms are subsequently presented to geologists and ontologists to assess their alignment with the domain's understanding (Sect. 3). The evaluation shows promising results. The entire corpus for this study is publicly available.

2. Methodology

In our pursuit of identifying terms relevant to the Geological Fault Domain, we adapted the methodology from Garcia et al. [18]. Our approach consisted of the following steps: (i) the selection of scientific journals within the domain of interest; (ii) the compilation of a comprehensive corpus comprising abstracts; (iii) the application of TF-IDF analysis to determine the primary keywords for the Geological Fault Domain; and (iv) a final evaluation of these keywords by domain experts against a pre-established ontology (Figure 1).

Figure 1: Methodology for identification of keywords for the Geological Fault Domain

TF-IDF (Term Frequency - Inverse Document Frequency) is a well-established method in information retrieval for evaluating the importance of keywords within a corpus. Essentially, it identifies terms that are frequent within a specific document but relatively rare across the entire corpus. This approach aids in highlighting words and expressions that best describe a particular document. Differently from Garcia et al. [18], we included in the analysis sets of documents with several levels of domain focuses, including general academic papers. If only one domain is analyzed, important expressions could seem common, when they are relatively rare compared to documents of different subjects. Using several degrees of focus allows us to highlight important terms and expressions for every domain.

The TF-IDF score for each term t in a document d is calculated as the product of two main components: Term Frequency (tf) and Inverse Document Frequency (idf). The term frequency, denoted as $tf(t,d)$, is the sum of term t occurrences in document d. In contrast, the inverse document frequency is given by the formula:

$$idf(t) = log\left[\frac{n}{df(t)}\right] + 1$$

Where:

- n represents the total number of documents in the document set.
- $df(t)$ is the number of documents in the document set that contain the term t.

Ultimately, the TF-IDF score for a term t in document d is computed as:

$$TF - IDF(t,d) = tf(t,d) * idf(t) = tf(t,d) * \left(log\left[\frac{n}{df(t)}\right] + 1\right)$$

For our TF-IDF calculation, we utilized abstracts of academic papers as our documents. Abstracts offer distinct advantages, as they condense essential information and vocabulary into concise sentences and are accessible for a wide range of papers, including those not available as open-access. Our initial step involved the selection of relevant scientific papers. The TF-IDF method compares the vocabulary of a specific set of texts against the broader corpus. To ensure a diverse vocabulary and to balance the focus and scope of fault-related terms, we initially chose scientific journals within the Geological Fault Domain. Gradually, we expanded our selection to

encompass papers from broader knowledge areas. Figure 2 illustrates the distribution of papers across different knowledge areas and highlights our chosen journals.

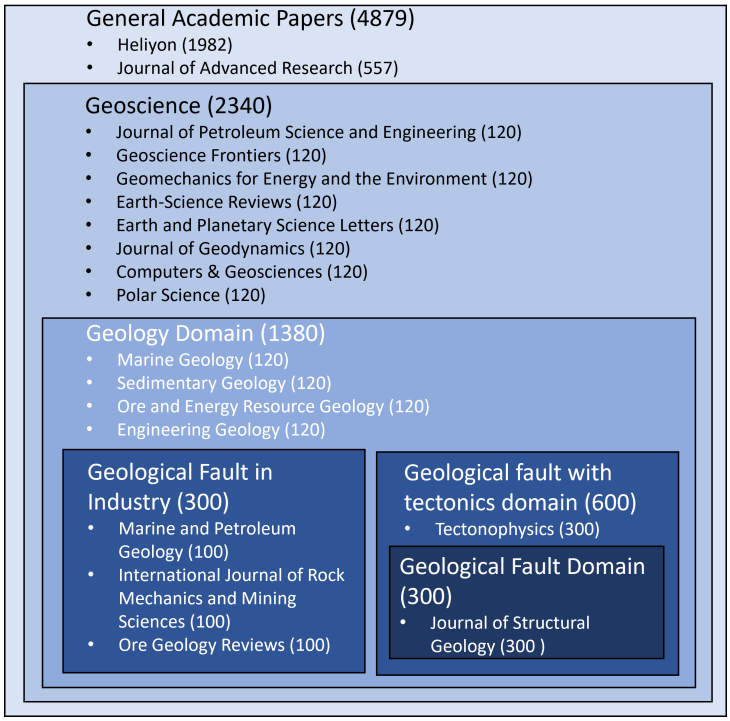

Figure 2: Distribution of paper by knowledge areas and scientific journals. The numbers in parenthesis are the total of abstract extracted.

Once we compiled the corpus containing all abstracts, we preprocessed it by removing stop words, applying stemming (a technique that reduces words to their root or base form), and converting all text to lowercase. Subsequently, we calculated the TF-IDF scores for all words and expressions across all documents, considering single words (1-gram) as well as two (2-gram) and three-word (3-gram) expressions. Summing the TF-IDF values across documents within our domain of interest resulted a list of the most significant keywords.

3. Results and Evaluation

Results. In our investigation, we compiled lists of the most important terms and expressions for all subdomains of Figure 2; Tables 1 and 2 show respectively the keywords for 'Geological fault domain' and 'Geological fault with tectonic domain'. We calculate the TF-IDF value for each word of the vocabulary (and for the 2-gram and 3-gram expressions) for every paper. Then, we summed and sorted all TF-IDF values of the documents of the same domain.

Besides the lists of important keywords for the geological fault domain, this paper presents two noteworthy outcomes:

1. We have assembled a corpus comprising 4,879 scientific papers that focus on various geoscience domains.

1-gram	Σ TF-IDF	2-gram	Σ TF-IDF	3-gram	Σ TF-IDF
fault	48.59	shear zone	5.58	strike slip fault	2.7
zone	16.29	strike slip	5.34	fold thrust belt	1.86
deform	15.24	fault zone	5.07	fault bend fold	1.29
fractur	15.16	normal fault	4.37	fault slip data	1.2
fold	14.57	damag zone	4.3	pull apart basin	1.14
slip	13.93	slip fault	2.93	fault propag fold	1.1
structur	13.83	fault core	2.88	fault damag zone	0.84
shear	12.3	fractur network	2.49	anisotropi magnet suscept	0.74
thrust	11.62	fold thrust	2.35	pre exist structur	0.61
strike	8.67	nw se	2.24	strike slip shear	0.6
stress	8.49	fault slip	2.12	dextral strike slip	0.57
model	8.23	fault rock	2.12	tecton shear stress	0.5
rock	8.08	thrust belt	2	fault stress regim	0.49
tecton	7.97	pre exist	2	thrust fault stress	0.49
strain	7.33	thrust fault	1.97	virtual outcrop model	0.49
basin	6.92	fault band	1.85	actual contact area	0.45
kinemat	6.79	fault segment	1.77	damag zone width	0.45
normal	6.55	deform band	1.74	dip normal fault	0.44
detach	6.29	fault propag	1.65	slip shear zone	0.44
seismic	6.15	ne sw	1.6	low angl fault	0.43

Table 1
List of the 60 most relevant terms after calculating and summing TF-IDF for all abstracts of "Geological Fault Domain" journals. They are split in single words term (1-gram), and expressions with two (2-gram) and three (3-gram) words (domain experts makes the less relevant terms in red color).

2. We have demonstrated a methodology for compiling documents and extracting important keywords from them.

Evaluations. In collaboration with geologists, we analysed the alignment between the high-frequency terms in our results and the domain knowledge. For the top-20 high-frequency terms in the Geological Fault Domain (table 1), 49 of 60 terms are closely related to the knowledge of fault; in the results of Geological Fault with tectonics Domain (table 2), only 40 of 60 terms are closely related to the knowledge of fault. It's worth noting that geologists identified some terms as "noisy," as they are used to describe faults or are related to specific study areas.

In addition to relevance checks, there are some interesting analysis results from the discussion between geologists and ontologists. In Tables 1 and 2, the terms *shear zone* and *fault zone*, in geologists' view, are interchangeable in the context of brittle deformation. The terms *damage zone* and *fault core* are two components of *fault zone*. The term *fault rock* shares a certain level of similarity with *fault core*, but not necessarily the same. The *pull apart basin* is the result of *normal fault*, and *thrust fault* is a type of *low angle fault*. These distinctions, while clear to geologists in an academic context, highlight potential semantic ambiguities in everyday usage. Such distinctions prove invaluable when geologists seek specific data and information from databases. Additionally, we also noticed that our data sources are academically biased, which contributes to the presence of certain noisy terms.

1-gram	Σ TF-IDF	2-gram	Σ TF-IDF	3-gram	Σ TF-IDF
fault	85.02	strike slip	10.34	strike slip fault	5.47
zone	27.33	fault zone	9.07	fold thrust belt	2.76
slip	27.01	normal fault	7.73	fault damag zone	1.48
deform	24.68	shear zone	7.33	pull apart basin	1.39
earthquak	24.21	slip fault	5.81	fault bend fold	1.34
structur	23.06	damag zone	5.03	fault slip data	1.26
seismic	21.05	nw se	3.5	fault propag fold	1.23
stress	18.62	fold thrust	3.45	fault stress regim	1.04
thrust	18.52	thrust belt	3.4	dextral strike slip	0.94
fold	18.12	ne sw	3.34	later strike slip	0.93
fractur	17.92	fault slip	3.24	philippin sea plate	0.93
shear	17.55	fault core	3.17	seismic reflect profil	0.92
strike	16.99	slip rate	3.15	anisotropi magnet suscept	0.91
tecton	16.29	thrust fault	3.08	strike slip motion	0.82
model	16	tibetan plateau	2.89	apatit fission track	0.82
basin	14.78	pre exist	2.87	normal fault earthquak	0.79
km	13.31	stress field	2.76	southern qilian shan	0.76
ruptur	13.24	fault segment	2.73	play import role	0.76
strain	12.9	strain rate	2.73	north china craton	0.75
region	12.15	fractur network	2.64	ne sw trend	0.74

Table 2
List of the 60 most relevant terms after calculating and summing TF-IDF for all abstracts of "Geological fault with tectonics" journals. They are split in single words term (1-gram), and expressions with two (2-gram) and three (3-gram) words (domain experts makes the less relevant terms in red color).

4. Conclusion

This paper proposes an approach with TF-IDF to quantify the term frequency of geological faults during the conceptualization phase of ontology development. The experiment has yielded interesting results for both geologists and ontologists. Our experiment contributes to developing basic terminology for creating knowledge models of geological faults, which will facilitate the retrieval of geological information and data from various sources. The experiment results also provide a basis with good knowledge coverage for ontological analysis to help geologists and ontologists deconstruct the semantically overloaded term 'fault' and its various hidden meanings. In future research, we plan to 1. refine the identified terms for improving the quality of the vocabulary list; 2. incorporate more industrial and technical documents for a more comprehensive analysis; 3. conduct ontological analyses of the terms and implement the proposed corpus to support the development of Petrobras' in-house search engine.

Acknowledgments This work was partially supported by the Norwegian Research Council via SIRIUS (237898), PeTWIN (294600) and Petrobras Researcher Center (CENPES).

Code availability: https://github.com/fabiocorreacordeiro/GeoscienceCorpus

References

[1] A. Torabi, H. Fossen, A. Braathen, Insight into petrophysical properties of deformed sandstone reservoirs, Aapg Bulletin 97 (2013) 619–637.

[2] Y. Qu, B. Zhou, E. Kharlamov, M. Giese, Industrial geological information capture with geostructure ontology, in: Proceedings of the 1st International Workshop on Semantic Industrial Information Modelling (SemIIM 2022) co-located with 19th Extended Semantic Web Conference (ESWC 2022), 2022.

[3] Y. Gil, C. H. David, I. Demir, B. T. Essawy, R. W. Fulweiler, J. L. Goodall, L. Karlstrom, H. Lee, H. J. Mills, J.-H. Oh, et al., Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance, Earth and Space Science 3 (2016) 388–415.

[4] Z. K. Shipton, J. J. Roberts, E. Comrie, Y. Kremer, R. Lunn, J. Caine, Fault fictions: Systematic biases in the conceptualization of fault-zone architecture, Geological Society, London, Special Publications 496 (2020) 125–143.

[5] Y. Qu, M. Perrin, A. Torabi, M. Abel, M. Giese, Geofault: A well-founded fault ontology for interoperability in geological modeling, arXiv preprint arXiv:2302.07059 (2023).

[6] N. Guarino, D. Oberle, S. Staab, What is an ontology?, Handbook on ontologies (2009) 1–17.

[7] L. F. Garcia, M. Abel, M. Perrin, R. dos Santos Alvarenga, The geocore ontology: a core ontology for general use in geology, Computers & Geosciences 135 (2020) 104387.

[8] J. Zhong, A. Aydina, D. L. McGuinness, Ontology of fractures, Journal of Structural Geology 31 (2009) 251–259.

[9] H. A. Babaie, A. Davarpanah, Semantic modeling of plastic deformation of polycrystalline rock, Computers & Geosciences 111 (2018) 213–222.

[10] A. Mantovani, F. Piana, V. Lombardo, Ontology-driven representation of knowledge for geological maps, Computers & Geosciences 139 (2020) 104446.

[11] S. J. Cox, S. Richard, A geologic timescale ontology and service, Earth Science Informatics 8 (2015) 5–19.

[12] F. CICCONETO, GeoReservoir: An ontology for deep-marine depositional system description, 2021. URL: https://lume.ufrgs.br/bitstream/handle/10183/220455/001124842.pdf?sequence=1&isAllowed=y.

[13] P. F. d. Silva, ResRiskOnto: an application ontology for risks in the petroleum reservoir domain, 2022. URL: https://www.maxwell.vrac.puc-rio.br/58981/58981.PDF.

[14] R. K. Romeu, F. C. Cordeiro, M. d. C. Rodrigues, D. d. S. M. Gomes, A. M. A. Alexandre, Busca semântica (tipo google) para recuperação mais inteligente de informação de reservatórios e exploração, ????

[15] Petrobras, PUC-Rio/ICA, Petrolês - corpus for the oil and gas industry., ???? URL: https://petroles.puc-rio.ai/index_en.html.

[16] N. F. Noy, D. L. McGuinness, et al., Ontology development 101: A guide to creating your first ontology, 2001.

[17] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernandez-Lopez, The neon methodology framework: A scenario-based methodology for ontology development, Applied ontology 10 (2015) 107–145.

[18] L. F. Garcia, F. H. Rodrigues, A. Lopes, R. d. S. A. Kuchle, M. Perrin, M. Abel, What geologists talk about: Towards a frequency-based ontological analysis of petroleum domain terms., in: ONTOBRAS, 2020, pp. 190–203.

[19] L. Shi, C. Jianping, X. Jie, Prospecting information extraction by text mining based on convolutional neural networks–a case study of the lala copper deposit, china, IEEE access 6 (2018) 52286–52297.

[20] E.-J. Holden, W. Liu, T. Horrocks, R. Wang, D. Wedge, P. Duuring, T. Beardsmore, Geodoca–fast analysis of geological content in mineral exploration reports: A text mining approach, Ore Geology Reviews 111 (2019) 102919.

[21] Y. Ma, Z. Xie, G. Li, K. Ma, Z. Huang, Q. Qiu, H. Liu, Text visualization for geological hazard documents via text mining and natural language processing, Earth Science Informatics (2022) 1–16.

[22] Q. Cao, X. Cheng, S. Liao, A comparison study of topic modeling based literature analysis by using full texts and abstracts of scientific articles: a case of COVID-19 research 41 (????) 543–569. URL: https://www.emerald.com/insight/content/doi/10.1108/LHT-03-2022-0144/full/html. doi:10.1108/LHT-03-2022-0144.