E-038

知識グラフを用いた社内情報検索システムの検討
Enterprise knowledge system with knowledge graph

齋藤 靖二† ホアン レアン† 佐野 恵一† 鈴木 健太郎†
Seiji Saito Hoang Le anh Sano Keiichi Kentaro Suzuki

1. はじめに

企業活動をしている会社内部には、業務遂行に必要なノウハウ・知識といった情報がサーバなどに保管されている。業務の規模が大きい場合、このような情報が膨大かつ複数のサーバに点在していることがあるため、情報を見つけられずに業務に時間をかけてしまい、機会損失に繋がることが考えられる。このため、ノウハウ・知識が記載された文書を登録し、検索によって早くたどり着ける社内検索プラットフォームは業務効率化とって非常に有用である[1]。一般的な検索手法として、全文検索が挙げられる。全文検索は検索対象となる文書に検索キーワードが含まれていない場合、検索結果に表示されない課題がある。

この課題を解決するために、文書を表す特徴語を抽出した後、文書と特徴語を紐づけることで知識グラフを生成し、共通の特徴語を介して関連する文書を抽出する方法を検討した。本稿では、知識グラフを活用した検索と全文検索とを比較することで本法の有効性の検証を行った。

2. 知識グラフを用いた検索システム

2.1 知識グラフを活用した検索支援の方法

全文検索では、検索キーワードに一致したワードを含む文書を検索結果として抽出し、必要な情報をユーザーに提供できる。しかしながら、検索キーワードに一致もしくは類似した単語を含んでいない場合、検索意図に関連する文書や必要な情報を含んだ文書であっても、検索結果としては抽出されない。この課題を解決するため、ユーザー自身が最適な検索ワードを思い付かない場合においても、特徴語を用いることで追加検索を進めることができる方法を検討した。

この検索に必要となる特徴語は、事前に生成した知識グラフから抽出した。図１に、品質不具合事例の文書をモチーフとした知識グラフの例を示す。“コネクタ破損”、“部品不良”、“配線不具合”という３つの文書を対象とし、共通の単語を含む文書同士が、特徴語により関係性の紐付けがされている。ここでは文書と特徴語がノードとなり、文書の中に特徴語を含むという関係をエッジとしている。

特徴語がそれぞれの文書から“クラック”、“ハンダ”、“非正規品”、“ハーネス”などが特徴語として抽出された場合、特徴語を含む文書と紐づけるため、この特徴語が媒介となって文書同士が関係性を持つ構造を構築できる。生成された知識グラフから、対象となる文書に紐づく特徴語を抽出し、ここからユーザーが希望する特徴語を選択することで関連する文書を追加で検索できる。これにより、全文検索だけでは、見つけられなかった文書も抽出可能となる。また、複数の特徴語を選択すると全てが紐づいた文書だけが抽出されるため、検索対象を絞ることもできる。

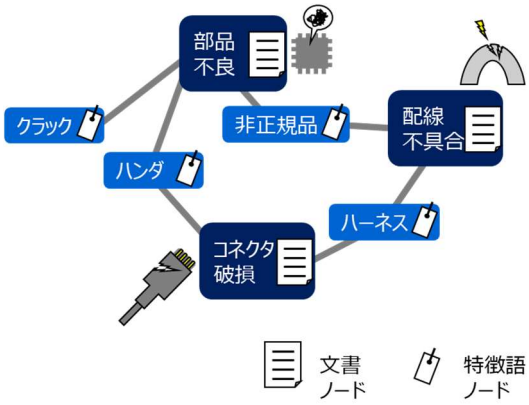

図 1 知識グラフの例

一方、１つの文書から複数の特徴語ノードとエッジが登録される特性上、データ量は膨大となる。このため、知識グラフを使った検索は、特定カテゴリが付いた事例もしくは規程などの特定領域の文書を対象とした。本稿では、社内で公開している約 300 件の特定書式で記載された品質不具合事例をモチーフとして手法の試行と検証を行った。

2.2 特徴語の抽出

知識グラフの生成の主な処理フローを図 2 に示す。入力された文書のテキストデータを形態素解析によって単名詞に分解し、単名詞が連続することで表される複合語も含め特徴語候補として抽出する。

特徴語候補抽出の後、FLR 法[2]を用いて重要度を算出し、この値が高い上位単語を特徴語として抽出する。なお、FLR は(1)の式で表される。

$$FLR(CN) = f(CN) \times LR(CN) \quad (1)$$

図 2 特徴語抽出フロー

†(株)東芝 Toshiba Corporation

ここで CN は特徴候補語、f(CN)は CN が単独で出現した頻度を表している。また、LR(CN)は FL(N)を単名詞 N に対する左側に連結する出現回数、FR(N)を単名詞 N に対する右側に連結する回出現回数、L を複合語の単名詞数を正規化するための変数とすると (2)式で表される。

$$LR(CN) = \left(\prod_{i=1}^{L}(FL(N_i)+1)\times(FR(N_i)+1)\right)^{\frac{1}{2L}} \quad (2)$$

抽出した特徴語の中には、品質不具合事例の書式に含まれる固定的な文章が含まれている。書式として共通的に出てくるワードは、どの文書に対しても出現するため、この文書を表す特徴語としては適さない。この判断にもとづき、該当のワードを、特徴語から削除する後処理を追加した。この処理を経て抽出した特徴語は、抽出元の文書と関連するノードとして登録した。

2.3 関連検索の機能検証

全文検索との比較で、複数の文書から作成された知識グラフを使った検索の効果を検証した。最初に、知識グラフを使った関連検索の流れを説明する。図 3 に関連検索処理の流れを、図 4 に検索画面のイメージを示す。検索画面上部の検索キーワード入力領域にフリーワードを入力し、全文検索を実行する。全文検索の結果、5 件の文書が表示された場合、この 5 つの文書に紐づけられた特徴語が表示される。この特徴語を選択することで、選ばれた全ての特徴語に紐づく文書のみが表示されていくため、絞込みワード候補として表示されている。図 4 では“クラック”と“検査”の 2 つの特徴語を選択することにより、3 つの文書数まで絞り込むことができている。最終的に、絞り込まれた

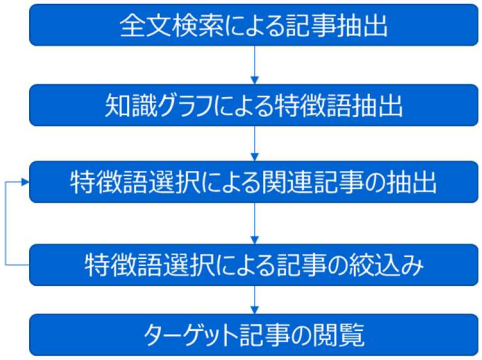

図 3 検索処理フロー

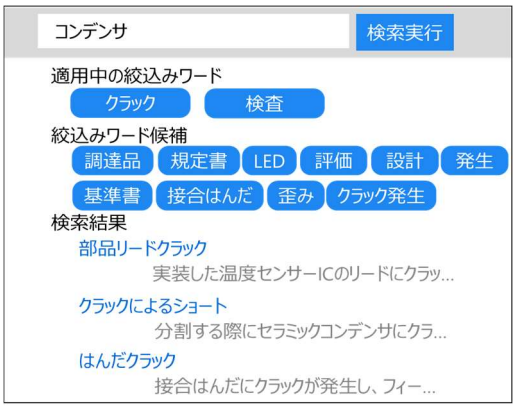

図 4 検索画面

表 1 全文検索と関連検索の検索結果文書数

ターゲットとした文書の内容	検索キーワード	抽出件数	
		全文検索	関連検索
クラックによるショート	コンデンサ クラック 検査	7	5
電源ユニット不具合	電源 故障	23	3
ベアリング不具合	振動 製造委託 調達	11	2

文書から閲覧したい文書を選択することで、文書の内容を閲覧することができる。

以上の流れで実施した関連検索と、本稿で文書を絞り込む際に選択した特徴語による AND の全文検索とを比較して検証を行った。表 1 に結果を示す。全文検索では検索キーワードに記載している特徴語で AND 検索を行った。関連検索では、最初に下線が引かれた検索キーワードを使って全文検索したのち、他の検索キーワードを用いて知識グラフから文書の絞込みを行った。

比較の結果、全文検索に比べ関連検索の文書数は少なく、表示された文書リストからターゲット文書を見つけ易いという傾向があった。これは、全文検索に比べ、各文書を特徴語に絞って検索を行うことの有効性を示していると考えられる。また、関連検索の場合は特徴語を提示されるため、検索キーワードが思いつかない場合でも検索することができる利点があるといえる。仮に、最適な検索キーワードを入力できた場合は、最初の全文検索で文書が見つかる可能性があるため、その後の関連検索は不要となる。このように最初の全文検索の結果によって関連検索の追加使用を使い分けることで、よりよい検索環境を提供できると考える。

以上の結果から、全文検索だけでは取得出来ない関連文書の抽出や、提示した特徴語を使った文書の絞込みが行える検索システムを実現した。これにより検索キーワードに一致する情報に加え、関連する特徴語を含んだ情報を検索でき、求めていた情報を効率的に取得できることを確認した。

3. おわりに

社内検索システムの全文検索に加え、知識グラフを用いた関連検索について検討を行った。社内検索システムのすべての文書を対象とした場合、関係情報が膨大になるため、領域を絞り適用することとした。最初のモチーフとして品質不具合事例の文書を取り上げ、知識グラフを使った関連検索と全文検索の比較を行った。今後、文書が追加された場合の知識グラフの更新など運用を考慮した課題を洗い出し、社内検索システムへの実装を進めていく。

参考文献

[1] 佐野恵一，“階層的データ管理と複数データ領域の高速横断検索を実現する社内ナレッジシステム”，FIT2022 O-021
[2] 中川ら，“出現頻度と連接頻度に基づく専門用語抽出”，自然言語処理 2003 年 10 巻 1 号 p. 27-45，(2003)