6C-02

業務知識を反映したオントロジーを用いた高精度な社内文書検索方式の実現

西出　恭平†　　阪田　恒次†

三菱電機株式会社　情報技術総合研究所†

1. 背景

「省エネ」に関する社内文書を検索する時、「省エネ」という用語が含まれる文書に加え、関連用語が含まれる文書を利用者に提示する必要がある。本研究では、オントロジーを活用して、検索文の関連用語を含む文書を検索できるようにすることを目指す。

2. 関連研究

社内文書には専門用語や業務知識が多く含まれている。これらをオントロジーで管理し検索システムに活用する研究がある。[1]では建築分野の社内文書を対象に、記載内容の共通点から分類を抽出し、分類別にオントロジーを構築して検索を試みている。

しかし、社内文書の内容やファイル形式が多岐にわたる場合、共通点がなく、分類別に構築したオントロジーを用いて検索ができない。多岐にわたるファイルを一括して検索するため、分類別のオントロジーを一つに集約すると、同一概念が複数の箇所に表れてしまい意図していない階層関係が構築されることがある。

3. オントロジーを活用した文書検索手法

3.1. 文書検索手法の要件と全体像

本研究では下記要件を満たす手法を検討する。

- 複数のオントロジーを集約して構築したオントロジーを用いて、関連用語を含む文書を検索できる。
- 集約で意図しない階層関係を構築しても利用者に適切な検索結果を提示するため、検索文の上位概念の用語は考慮せず、下位概念に含まれる用語のみ考慮して検索できる。

「LED」と検索したときに、利用者は LED の消費電力や型番など、具体的な内容の関連情報を求める。しかし上位概念を参照すると、「照明」のように LED よりも抽象度が高い用語を検索文に含めるため、意図から外れた結果が表示される。そのため下位概念のみを参照し、LED を具体化する用語を含めて検索をすることで、

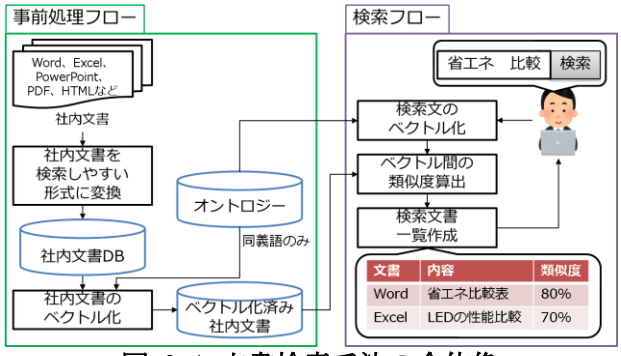

図 3-1 文書検索手法の全体像

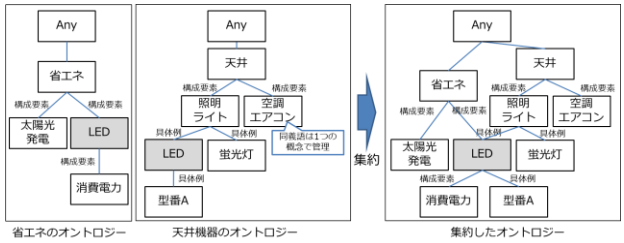

図 3-2 オントロジー集約

利用者の意図に沿った検索を実現する。

文書検索手法の全体像を図 3-1示す。文書検索手法では、ベクトルに変換した社内文書と検索文の類似度をもとに提示する文書を決定する。検索文の下位概念のみ考慮した検索をするため、用語間の重みは検索文にのみ加算する。

3.2. オントロジーの構築

オントロジーは、「概念」と概念間の「関係」を階層的に表した、人の知識をコンピューターに理解させる表現方法の一つである。概念間の階層が近いものは意味が近く、離れているものは意味が遠いこと表す。

複数のオントロジーを一つに集約した例を図 3-2に示す。「Any」は最上位を表す概念である。「空調」と「エアコン」のように同義語は同じ概念で管理する。また、「LED」は省エネと天井機器のオントロジーの両方に含まれる概念であり、集約すると、「LED」は「省エネ」と「照明」の二つの上位概念と、「消費電力」「型番A」の二つの下位概念を持つ。

3.3. ベクトル構築処理

検索文と社内文書のベクトル化処理には、オ

Internal document retrieval method with ontology reflected domain knowledge
†Kyohei Nishide, Koji Sakata
Information Technology R&D Center, Mitsubishi Electric Corporation

表 3-1 オントロジーの重みの算出例

		参照先概念						
		Any	省エネ	太陽光発電	LED	消費電力	型番 A	天井
参照元概念	Any	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	省エネ	0.00	0.00	0.92	0.95	0.87	0.87	0.00
	太陽光発電	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LED	0.00	0.00	0.00	0.00	0.99	0.99	0.00
	消費電力	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	型番 A	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	天井	0.00	0.00	0.00	0.90	0.84	0.84	0.00

ントロジーの上位下位関係の反映が容易な BoW（Bag of Words）と TF-IDF を用いる。BoW では形態素解析の結果、品詞が名詞の単語のみを用いてベクトルを構築した。

さらに、検索文の関連用語を含む文書を検索するため、検索文にオントロジーの用語間の重みを加える。オントロジーの用語間の重みの算出方法とベクトルへの重みづけ方法は、[2]に記載された方法を用いる。ただし、[2]の方法では、上位概念や兄弟の位置にある概念に対しても重みが付けられてしまう。そのため、ある概念の下位に含まれない概念への重みは 0 とする。

表 3-1に図 3-2のオントロジーの重みを一部示す。「省エネ」と検索した時は、下位概念である「太陽光発電」と「LED」、さらに「LED」の下位概念の「消費電力」と「型番 A」の重みを考慮した検索文ベクトルを構築する。「太陽光発電」と検索した時は他の用語への重みはすべて 0 であるため、上位概念の用語は考慮しない。

3.4. 類似度の算出と検索文書の一覧作成処理

社内文書のベクトルと検索文のベクトルの類似度はコサイン類似度を用いる。検索文書の一覧は、類似度を降順に並べ替えたものである。

4. 評価

4.1. 評価目的と評価方法

オントロジーの用語間の重みを考慮することで、文書の検索精度が向上するのか明らかにすることを目的に評価した。

評価は 1249 文書を対象に、25 名で作成した 223 件の検索事例（検索文と目的文書の組）を用いて、検索文を入力した時に目的文書が上位に表れたか確認した。評価指標には平均逆順位を用いる。あわせて目的文書が上位 20 件以内に表れた割合を算出する。比較手法には、オントロジーの利用方法に①形態素解析の辞書に用語を追加登録のみと②同義語と上下概念への重みを考慮した手法を用いた。なお、比較手法②において用語間の重み算出の際に、最上位の要素である「Any」が共通の親概念の場合は、別分類のオントロジーであると判断し、用語間の重みは 0 とした。また評価には、158 用語を含むオントロジーを用いた。

表 4-1 評価結果

オントロジーの利用方法	[提案手法] 同義語と下位概念への重み考慮	[比較手法①] 形態素解析の辞書に用語を追加登録のみ	[比較手法②] 同義語と上下概念の重み考慮
平均逆順位	0.48	0.43	0.43
上位 20 件に目的文書の出現割合	82.1%	75.8%	82.1%

4.2. 評価結果

評価結果を表 4-1に示す。提案手法の平均逆順位が最も高いため、複数のオントロジーを集約しても、下位概念のみを考慮することで高精度に検索できることが分かった。また、オントロジー上位下位概念を考慮するよりも、下位概念のみを考慮したほうが、平均逆順位が高かったため、より利用者の意図に適した検索結果を提示できたと判断する。

上位に出現しなかった検索事例には、「省エネ　比較　条件」のように複数の単語で検索をした時に、「省エネ」と「比較」は TF-IDF 値が高く、「条件」は低いため、「省エネ　比較」と類似した文書が上位に出現していた。目的文書が検索文の用語をすべて含む場合に、検索順位が低いことが見られたため、今後はこの課題への対策を実施する。

5. おわりに

オントロジーを活用した文書検索手法を報告した。今後は、精度向上策を検討するとともに、オントロジーの管理・構築を容易にする方法を検討する。

参考文献

[1] 古川慧他「建築分野における社内技術文書検索システム作成のための予備的調査」人工知能学会全国大会論文集, 2020

[2] Liping Jing,et al."Knowledge-based vector space model for text clustering", Knowledge and Information Systems Vol.25 Issue 1, pp.35-55, 2010