文書検索の精度を向上させる文書分割アルゴリズム

総合情報学部 総合情報学科
安達 由洋 教授 Yoshihiro Adachi

研究概要　意味分散表現に基づいて文書をそれが含む主題に分割し索引付けして検索することで、文書検索の精度を向上させる文書分割アルゴリズムを開発しました。

研究シーズの内容

文書データの検索には伝統的にキーワードによる検索手法が使用されてきました。また、文書分類・検索のための索引付けとしては単語の出現頻度に基づくTF-IDTが使用されてきました。しかし近年、単語や文章などの自然言語が持つ意味情報を分散(ベクトル)表現するSkip-gram、PV-DM、SCDVなどの技術が登場しています。文書データをその意味分散表現ベクトルで索引付けすると、正確なキーワードが分からなくても検索したい意味や話題を表す質問文を入力すれば、その意味や話題を含む文書データを検索することができ、知的で柔軟な文書検索システムが実現できます。

ただし、新聞記事や雑誌などの多数の主題を含む文書に対しては、従来の研究で用いられている文書全体の内容情報(意味情報あるいはTF-IDTなど)による分散表現では、個々の主題による特徴を平均した索引付けとなり、特定の主題に対する検索を行ったとき検索精度が低くなります。

この課題を解決するため、文書を先頭から次々と断片に切り出し、その断片の分散表現から主題(内容情報)の変化を検出することで、多数の主題からなる文書を主題ごとに自動的に分割する文書分割アルゴリズムを開発しました。

このアルゴリズムでは、多数の主題を含む文書に対して、意味内容情報の分散表現に基づいて主題ごとに文書を分割することで、分割された文書断片の索引付けを用いて文書を検索することができ、精度良く検索したい内容を含む文書を検索できるようになります。

また、分割された文書断片に元の文書の対応する章やページ情報などをタグ付けすると、文書のどの部分(章、節、ページ)に検索対象の内容が書かれているかを提示できるようになります。

		Q1	Q2	Q3	Q4	Q5	平均
従来の分散表現手法の各検索評価値	Skip-gram	11.5	10	6	11.5	13.5	10.5
	CBOW	11	4	3	12	12.5	8.5
	PV-DM	7.5	5	0	11	7	6.5
	PV-DBOW	6	1	0	5	13	5
	TF-IDF	12	10	5	7.5	10	8.9
	SCDV(Skip-gram)	8	12	12	9.5	14	11.1
	SCDV(CBOW)	11	13	8.5	8	12.5	10.6
	non-SCDV	12	10	12	11	13	11.6
本手法	**non-SCDV**	**11**	**15**	**10**	**13**	**13.5**	**12.5**

表 1. 各分散表現手法と本手法の検索評価値比較

研究シーズの応用例・産業界へのアピールポイント

図書館レファレンスサービス、書籍販売、ネット上の文書検索など、文書検索分野全般に適用可能

特記事項(関連する発表論文・特許名称・出願番号等)

「レファレンスサービス自動化のための書籍分散表現」、FIT2019、情報処理学会.

特願 2019-148388、「部分単語列を生成する方法、部分単語列生成装置、部分単語列生成プログラム」.