N-034

施設配置モデルを用いた特許文献の要約文抽出手法の提案
A facility location approach to patent document summarization

矢野裕和 † **古田壮宏** ‡ **赤倉貴子** ‡
Hirokazu Yano Takehiro Furuta Takako Akakura

1. はじめに

近年，知的財産への関心の高まりから，非常に多くの特許出願がなされている．特許出願後，原則 1 年 6ヶ月を経過すると公開特許公報が発行される．公開特許公報は技術文書としての役割があり，特許調査などに用いられている．しかし，年間 30 万件以上の公開特許公報が発行されているため [1]，必要な情報に辿り着くのは困難となっている．

大量の公開特許公報の中から必要とする情報を探し出すためには，発明の概要を迅速かつ的確に把握する必要がある．このためには，特許出願に必要な書類の一つである要約書の利用が考えられる．要約書は，発明が解決しようとする課題とその解決手段で構成されている．

しかし，特許出願書類の記述量に関係なく，400 字以内で記述しなければならないという規則がある．また，要約書の解決手段で書かれている内容は，請求項そのものであることが多いため，請求項を読んで得られる情報とあまり変わらず，要約書を読むメリットが少ないと考えられる．特に，請求項は 1 文で記述されるという独自の記述スタイルをとるため，非常に読みづらくなっている．よって，特許出願に添付されている要約書は発明の内容の把握に十分な情報を含んでいないと考えられる．そこで，請求項の内容を詳しく説明している明細書を用いて要約書を作成することで，発明が理解しやすくなると考えられる．

本稿では，明細書中の特徴を考慮した要約文抽出を目指し，既存の重要文抽出手法を公開特許公報に適用することを試みる．高村らは，一般の文書要約を対象として，オペレーションズリサーチにおける施設配置問題に基づく重要文抽出手法を提案している [2]．本稿では，高村らの手法を基に，明細書からの要約文抽出手法を提案する．その上で，実際の公開特許公報を用いて提案手法の適用可能性を示す．

2. 施設配置問題による文書要約のモデル化

文書要約の代表的な手法として，与えられた文章から必要な文を選択して要約を生成する文選択がある [3]．この手法は，与えられた文章が文法的であれば，選択された文も文法性は保障されているという特徴があり [2]，本稿でもこのような要約を目指して，可能な限り元の文章の内容を多く含む重要な文を抽出することを行う．明細書のすべての文章が，選択された明細書の文集合のいずれかの文によってできる限り表現されるような要約モデルを考える．本稿では，施設配置問題（p-median problem [4]）による文書要約のモデル化 [2] に基づき，公開特許公報の要約文抽出手法を提案する．

パラメータ

S ：明細書に含まれる文の集合
e_{ij} ：文間係数，文 $i \in S$ と $j \in S$ の含意関係
p ：抽出する文の数

決定変数

x_i ：文 $i \in S$ を重要文として抽出するか否かを表す 0-1 変数
z_{ij} ：文 $j \in S$ と共通する情報を多く含む文として文 $i \in S$ に割り当てる場合 1，そうでなければ 0 を表す 0-1 変数

これらのパラメータを用いて，重要文抽出モデルを以下のように定義する．

$$\max. \quad \sum_{i \in S} \sum_{j \in S} e_{ij} z_{ij} \tag{1}$$

$$\text{s. t.} \quad z_{ij} \leq x_i, \qquad i, j \in S \tag{2}$$

$$\sum_{i \in S} x_i \leq p \tag{3}$$

$$\sum_{i \in S} z_{ij} = 1, \qquad j \in S \tag{4}$$

$$x_i \in \{0, 1\}, \qquad i \in S \tag{5}$$

$$z_{ij} \in \{0, 1\}, \qquad i, j \in S \tag{6}$$

目的関数 (1) は，明細書の文章が抽出された p 個の文によって表現されている度合いを表し，これを最大化する．式 (2) は，各文 j は重要文として選択された文 i にのみ割り当てることができることを示している．式 (3) は抽出する文の数が p 個以下であることを規定している．また，各文 j は必ず抽出される文のいずれかに割り当てられることを式 (4) により保証している．式 (5) および (6) は標準的な 0-1 制約である．

ここで，文間の含意関係を表すスコアである文間係数 e_{ij} を算出する方法について述べる．本稿では，高村らの研究と同様に，各文同士の含意関係は，そこに含まれる単語がどの程度共通しているかによって，表されると見なす．よって，文 $i \in S$ に対する文 $j \in S$ の文間係数 e_{ij} は，文 j に含まれる単語のうち，文 $i \in S$ に含まれる単語と共通するものがいくつあるかで表し，以下のように定義する．

$$e_{ij} = \frac{|W(i) \cap W(j)|}{|W(j)|} \tag{7}$$

ここで，$W(i)$ は文 i に含まれる単語の集合とする．なお，本稿では，名詞，動詞および形容詞を対象とした．

†東京理科大学大学院工学研究科
‡東京理科大学工学部

表 1: 提案手法により抽出された要約文

	【0054】	【0097-3】	【0107】	【0116】	【0119-3】
文の内容	形態 1	形態 2	形態 3	請求項 1	効果
p=1	0	0	0	1	0
p=3	0	0	1	1	1
p=5	1	1	1	1	1

形態：実施形態, 効果：発明の効果

【請求項 1】
固体撮像装置であって、行列状に配置され、入射光の強度に応じた信号電圧を出力する複数の画素回路と、列毎に1つ設けられ、対応する列に配置された複数の画素回路により前記信号電圧が出力される複数の垂直信号線と、列毎に1つ設けられ、対応する列に設けられた前記垂直信号線に接続される複数のスイッチ部と、列毎に1つ設けられ、対応する列に設けられた前記スイッチ部を介して前記垂直信号線に接続され、前記垂直信号線に出力された前記信号電圧を増幅する複数の列増幅部とを備え、前記複数の画素回路の各々は、入射光の強度に応じた信号電荷を蓄積する受光部と、フローティングディフュージョンと、前記受光部と前記フローティングディフュージョンとの間に接続された転送トランジスタと、前記フローティングディフュージョンの電圧に応じた前記信号電圧を前記垂直信号線に出力する増幅トランジスタとを備え、前記固体撮像装置は、さらに、前記受光部から前記フローティングディフュージョンに信号電荷を転送する転送期間において前記スイッチ部をオフする制御部を備える固体撮像装置。

図 1: 請求項 1 の内容 (特開 2011-114731)

【0116】
（まとめ）以上より、本発明の第1から第3の実施形態に係る固体撮像装置は、行列状に配置され、入射光の強度に応じた信号電圧を出力する複数の画素回路111と、列毎に1つ設けられ、対応する列に配置された複数の画素回路により信号電圧が出力される複数の垂直信号線103と、列毎に1つ設けられ、対応する列に設けられた垂直信号線に接続される複数のスイッチ部（スイッチトランジスタ114）と、列毎に1つ設けられ、対応する列に設けられたスイッチ部を介して垂直信号線に接続され、垂直信号線に出力された信号電圧を増幅する複数の列増幅部115とを備え、複数の画素回路の各々は、入射光の強度に応じた信号電荷を蓄積する受光部（PD1）と、フローティングディフュージョン（FD1）と、受光部とフローティングディフュージョンとの間に接続された転送トランジスタ（NM1）と、フローティングディフュージョンの電圧に応じた信号電圧を垂直信号線に出力する増幅トランジスタ（NM3）とを備え、固体撮像装置は、さらに、受光部からフローティングディフュージョンに信号電荷を転送する転送期間（t4～t5）においてスイッチ部をオフする制御部110を備えている。

図 2: 抽出された【0116】(特開 2011-114731)

3. 公開特許文献を用いた評価実験

提案手法を用いて，実際の公開特許公報の明細書から要約文抽出を行った．抽出する文の数 p を変化させて，抽出される文の内容を確認した．さらに，重要文抽出手法として利用されている tf-idf による重要文を用いて [5]，提案手法による要約文との比較を行う．e_{ij} を求めるときと同一の形態素を用いて tf，idf を計算する．

$$tf_k = \frac{\text{文 } i \text{ での単語 } k \text{ の出現回数}}{\text{文 } i \text{ で出現する総単語数}} \tag{8}$$

$$idf_k = \log\left(\frac{\text{明細書の全文数}}{\text{単語 } k \text{ を含む文数}}\right) \tag{9}$$

$$\text{文 } i \text{ の重要度} = \frac{\sum_{k \in W(i)}(tf_k \times idf_k)}{|W(i)|} \tag{10}$$

式 (10) より重要度を求め，重要度の高い文から順に p 文抽出する．表 1 は公開特許公報 (明細書全文 250 文) に対して提案手法を適用した結果である．値が 1 のときにその文が抽出されたことを表している.【0054】は明細書の段落番号であり,【0054】の 1 文を抽出したことを表している．2 文以上で構成されている段落の場合は,【0097-3】のようにその段落の何文目を抽出したかを表している.【0116】は抽出する文の数 p の全てにおいて抽出されたことを表しており，文の内容は請求項 1 についての記述であることを表している．

図 1 は請求項 1 の内容で，図 2 は提案手法で抽出された【0116】の内容である．この 2 つの内容を比較してみると,【0116】の内容は請求項 1 の内容を表していることがわかる．また,【0116】は出願に添付されている要約書の内容を含んでいるものであった．また，この公開特許公報には 3 つの実施形態があるが，表 1 より，選択された文の内容は実施形態 1，2，3 それぞれについて説明している文を選択していることがわかる．

表 2: tf-idf による重要文

重要文	重要度	文の内容
【0017】	2.01	課題解決手段
【0018】	1.66	課題解決手段
【0048-1】	1.24	実施形態 1
【0084】	1.15	実施形態 2
【0120-1】	1.02	スイッチ部の説明

表 3: tf-idf による重要文

重要文	実際の明細書の文
【0017】	また、前記スイッチ部はMOSトランジスタであってもよい。
【0018】	また、前記MOSトランジスタのゲートアスペクト比は、前記増幅トランジスタのゲートアスペクト比以下であってもよい。
【0048-1】	列電流源112は、電流源トランジスタNM4を含む。
【0084】	この電圧供給回路201は、プルアップトランジスタPM12を含む。
【0120】	また、スイッチ部はMOSトランジスタであり、MOSトランジスタのゲートアスペクト比は、増幅トランジスタのゲートアスペクト比以下である。

表 4: 明細書全体の網羅率

	提案手法	tf-idf
p=1	0.090	0.004
p=3	0.138	0.018
p=5	0.184	0.026

よって，発明の内容を網羅した文が選択されたと考えられる．

表 2 は公開特許公報に対して tf-idf を計算して，重要文を求めた結果である．表 3 は tf-idf による重要文の一覧である．提案手法で抽出した文よりも短いものが多く，補足的な説明が多い文が重要文になっている．そのため，発明の概要を把握できる文としては不十分であると考えられる．

ここで，式 (11) より，提案手法による要約文と tf-idf による重要文について，明細書全体に対してどの程度網羅できているかを比較する．網羅率が高ければ，明細書全体の内容を含んでいる文が抽出できたと考えられる．

$$\text{網羅率} = \frac{\text{要約文/重要文の単語異なり数}}{\text{明細書全文の単語異なり数}} \tag{11}$$

表 4 より，提案手法の方が網羅率は高いことがわかる．今回は全文 250 文に対して 5 文選択したが，選択する文を増やせば網羅率が高くなると思われる．

4. まとめ

既存の重要文抽出手法を基に，要約文抽出方法を検討し，明細書全体を網羅した要約文を抽出できる可能性を示した．今後は請求項の情報を文間係数に取り入れて，より請求項の内容を網羅している文を抽出可能にしたい．また，網羅率が高く，要約文数が多くなりすぎないような抽出する文の数 p の決定が必要である．

参考文献

[1] 特許庁，http://www.jpo.jp/indexj.htm，参照 June，2011.

[2] 高村大地，奥村学，“施設配置問題による文書要約のモデル化，” 人工知能学会論文誌，**25**(1), pp.174–182，2010.

[3] 奥村学，難波英嗣，植田禎子，“自動要約，” 共立出版，2003.

[4] S. L. Hakimi , “Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph,” Operations Research, **12**(3), pp.450–450, 1964.

[5] 唯野良介，嶋田和孝，遠藤勉，“アスペクトごとの文の重要度と類似性判断に基づく複数レビューの要約，” 言語処理学会，第 16 回年次大会発表論文集，pp.587–590，PA2-29，2010.