検索結果の絞り込みのため語集合の整数線形計画ソルバを使用した特定

手島 亮太[†]　岡部正幸[††]　梅村恭司[‡]

豊橋技術科学大学大学院情報・知能工学専攻[†]

豊橋技術科学大学情報メディア基盤センター[††]　豊橋技術科学大学情報・知能工学系[‡]

teshima@ss.cs.tut.ac.jp, okabe@imc.tut.ac.jp, umemura@tut.jp

概要　キーワード検索では多くの結果が得られてしまうことで，必要としている情報を得るのが難しくなることがある．このようなとき，ユーザは結果を絞り込むために語を追加して続けて検索を行うことが考えられる．本稿では連続した検索において効果的である語を特定する方法として，整数計画問題を解くことで語の特定を行う方法を提案する．さらに，100 個の語を特定したときに元の検索結果をどれだけ被覆するかの実験を行い，従来手法より多くの検索結果を継続して検索できるようになったことを確認した．

キーワード　情報検索，サジェスト，整数計画問題

1 はじめに

ある話題についてキーワード検索を利用した結果として，一度に読み切れない数の検索結果が得られるという問題がしばしば起こる．こうした場合，検索結果の数を絞り込むために，追加の語を検索語に加えることが良く行われるが，効果的な語を考えるという作業は話題の内容に関する知識を必要とする作業である．検索結果の全体を俯瞰し，効果的な語を容易に特定できれば良いが，そのためには検索結果について読み解く必要があり，人間にはコストの高い作業である．このコストを軽減するために，検索結果の中から絞り込みに効果のある語の集合を自動特定するという問題に取り組む．

効果的な絞り込み語の集合を特定するために，従来の研究では貪欲法による特定を行っていた[1]．本稿では語集合の特定方法として，0-1 整数計画問題を解く手法を使うことで従来手法より多くの検索結果を連続した検索の候補として利用できるようになったことを報告する．

2 語集合の特定

システムが特定した語の候補から絞り込みのための語集合を決めるにあたり文献[1]の 2 つの指針を準用する．これは「語を追加しても候補が十分に残る語を特定する」という指針と「語の集合を無駄に大きくしないために，出現の相関の大きい 2 つの語があった場合はどちらか一方を選ぶ」という 2 つの指針である．本節ではそれに関係して従来手法及び提案手法について述べる．

2.1 従来の語集合の特定

先に述べた前者の指針と後者の指針は相反する性質を持つ．この 2 つの指針を両立させる方法として，従来の手法では貪欲法による特定を行った．貪欲法は各地点において最良の結果を選ぶという方法である．これは前者の指針を満たすと同時に，出現相関の大きい語が既に選ばれている語は，評価値が高くなりにくいことで選ばれにくくなるため，後者の指針を満たすことにつながる．

2.2 整数計画法による語集合の特定（提案手法）

絞り込みのための語集合によって多くの検索結果を継続して検索できるようにするという問題は，絞り込みの語によってより多くの検索結果を被覆する集合被覆問題と見なせる．したがって，本稿ではこの問題を 0-1 整数計画問題に定式化し，最適解を求めることで絞り込みのための語集合の特定を行うことにする．

検索結果集合を**M**，各単語を表す添字の集合を**N**で表す．このとき語集合の特定は，検索結果 i で単語 j が出現する場合 1，しない場合 0 となるa_{ij}，単語 j が絞り込みの語として選ばれる場合 1，選ばれない場合 0 となるx_j，選ばれた絞り込みの語で検索結果 i が被覆される場合 1，されない場合 0 となるd_iを使って次のように定式化できる．なお，κは特定する語の上限数を表す定数である．

表 1 擬似的な検索集合と100の絞り込みの語による被覆結果

検索クエリ	検索結果数	被覆数(IP)	被覆数(従来)	被覆率(IP)[%]	IP 対 従来手法
アルゴリズム	11747	3571	3569	30.3	1.000560381
クラスタリング	561	514	509	91.6	1.009823183
コンパイラ	767	708	700	92.3	1.011428571
圧縮	9384	3062	3061	32.6	1.000326691
雑音	4757	2254	2252	47.3	1.000888099
最適化	5534	2497	2495	45.1	1.000801603
パーソナルコンピューター	833	689	684	82.7	1.007309942
ロボット	2950	1744	1743	59.1	1.000573723
ソフトウェア	5007	2234	2233	44.6	1.000447828

$$\text{Maximize} \quad z = \sum_{i \in \boldsymbol{M}} d_i$$

$$\text{subject to} \quad \sum_{j \in \boldsymbol{N}} x_j \leq \kappa$$

$$\sum_{j \in \boldsymbol{N}} a_{ij} x_j \geq d_i \quad \forall i$$

3 評価

本稿では絞り込みの語について，どれだけの検索結果を継続して検索できるか及びどれだけ人が正しく読むことができるかの評価を行う．後者は形態素解析などを使わない本システムにおいて，特定される絞り込みの語をユーザがどれだけ判断できるかを示すための評価である．なお，システムの評価は NTCIR-1[2]のテストコレクションを利用して作った擬似的な検索結果集合を利用した．NTCIR-1 は論文の抄録を集めた情報検索のコレクションであり，擬似的な検索結果集合は特定の話題を表現する文字列を含む検索結果の集合である．

3.1 検索結果の被覆率

NTCIR-1 からいくつかの検索クエリによって得られた擬似的な検索結果集合と，従来手法及び整数計画法でそれぞれ100個の絞り込みの語を選んで検索結果をどれだけ被覆するかについてまとめたものが表1である．表の被覆数は，検索結果に絞り込みの語を追加して検索した際に継続して検索できる検索結果の数を表しており，全ての擬似検索集合で被覆数が増加していることが確認できる．

3.2 絞り込みの語集合の可読性

検索クエリ「ロボット」による検索結果集合において，整数計画法を使って 100 個の絞り込みの語を特定した結果が表2である．ここで正しい絞り込みの語は人間が正しく読み取れる語であり，誤った絞り込みの語はそうでない語のことを表している．

表 2 の結果を見ると「二足歩行ロボット」というロボットの種類を示す語や，「軌道制御」や「協調作業」などのロボットの動作を示す語などが得られていることがわかる．これらはロボットが持つ話題として検索の絞り込みを行うのに自然な語である．

表 2 検索集合「ロボット」から特定された絞り込みの語

正しい絞り込みの語 (93[%] = 93/100)	誤った絞り込みの語 (7[%] = 7/100)
制御方式　遺伝的アルゴリズム	^<
視覚情報　軌道制御	送システム
ナビゲーション　協調作業	アームによ
二足歩行ロボット　プロセス	ラクション
作業空間　対象物体	ator
並列処理　組立作業	実ロボット
評価関数　フレキシブルアーム	ンインターフェース
インタフェース　インピーダンス	
制御ユニット　演算遅れ時間	
姿勢制御　自己組織化	

4 結論

本稿では，語の特定のための指針に沿って 0-1 整数計画問題の定式化を行い，解を求める方法で絞り込みのための語の特定を行った．これにより，これまでの研究手法と比べてより多くの検索対象を続けて検索することが出来るようになった．また，さらに整数計画法では最適解が得られることから，本システムの語集合による検索結果の被覆について上界を示すことができた．

参考文献

[1] 手島亮太，岡部正幸，梅村恭司：検索結果の絞り込みのために有用な語集合の特定，第 20 回言語処理学会発表論文集，pp. 137-140, 2014.

[2] 神門典子，栗山和子，野末俊比古ほか：”NTCIR-1：情報検索システム評価用テストコレクション構築の方針と実際”，情報処理学会研究報告．情報学基礎研究会報告，99(20)，pp. 33-40, 1999.