ラフ集合を用いた情報検索手法の発展

船越 要

北陸先端科学技術大学院大学 情報科学研究科

1997 年 2 月 14 日

キーワード: 情報検索システム, ラフ集合, tolerance relation.

情報検索システム

今日数多くの文献情報が電子的に提供されており、しかもその量は爆発的に増大している。情報検索システムは、データベース中から利用者の興味に沿って文献を選択することを目的とするが、文献数の増大に伴い、検索効率向上の重要性は高まっている。

情報検索システムは $\mathcal{S} = (\mathcal{T}, \mathcal{D}, \mathcal{Q}, \alpha)$ と定式化できる。ここで $\mathcal{T}$ は語の集合、$\mathcal{D}$ は文献の集合 (各文献は $\mathcal{T}$ の部分集合)、$\mathcal{Q}$ は質問の集合 (各質問は $\mathcal{T}$ の部分集合)、$\alpha : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}^+$ は文献と質問の適合度を示す関数である。利用者が質問 Q を入力すると、システム $\mathcal{S}$ は文献を $\mathcal{D}$ の中から Q に沿うように α の値に従って提供する。

情報検索システムの概念モデル (文献を選択する手法) としては、ブール演算を用いたブール型モデルが主流である。これはブール型モデルが単純であり、システムに適用し易い為であるが、実際にはブール型モデルによって利用者の満足出来る検索結果を提供するのは困難である。このため、非ブール型のモデルによる情報検索手法が幾つか研究されている。知的情報検索と呼ばれる非ブール型モデルでは、語の示す概念の意味的な計算が行なわれ、ラフ集合を用いた検索システムもこの延長上にあると考えられる。

ラフ集合と Tolerance Relations

ラフ集合 (Rough Sets) は Pawlak によって 1980 年代前半に提唱された集合論の一種であり、曖昧、不確実な情報の取扱に優れている。ラフ集合の理解に必要な基本概念は、approximation space と lower/upper approximations である。全集合 (universe) 内の要素は equivalence relation と呼ばれる等値関係に従って equivalence class と呼ばれるグループに分類され、一つの equivalence class に含まれる要素同士は同定不能 (indiscernible) とさ

れる。equivalence class に分割された集合を approximation space と呼ぶ。approximation space 内の任意の部分集合 X は、X に包含されている equivalence classes の和集合 (lower approximation)、及び X との共有要素が空でない equivalence classes の和集合 (upper approximation) によって表現される。

ここで、equivalence relation は、反射律、対象律、推移律を満たすことが求められるが、情報検索の分野では、扱う語間の関係が推移律を満たすことは稀である。そのため、equivalence relation は強過ぎる関係となっている。例えば、ラフ集合の情報検索への応用は Srinivasan などによって 1980 年代後半に研究されているが、equivalence relation を使用しているため、これらのシステムでは完全な同義語 ("artificial intelligence" と "AI" のような) のグループ化に留まり、知的情報検索には至っていない。

推移律の制約を持たない tolerance relation と呼ばれる等値関係を用いたラフ集合が、1994 年に Skowron and Stephaniuk などによって提唱された。tolerance relation によって集合 U 内の要素は tolerance class と呼ばれるグループに分類されるが、equivalence class の場合と異なり、tolerance class 間では要素の共有が許容されている。tolerance class によって分類された集合を tolerance space と呼び、$\mathcal{R} = (U, I, \nu, P)$ と記述する。ここで U は全集合、 $I : U \rightarrow \mathcal{P}(U)$ はある要素に対する tolerance class を決定する関数、$\nu : \mathcal{P}(U) \times \mathcal{P}(U) \rightarrow [0, 1]$ は 2 つの集合の包含度を示す関数、$P : I(U) \rightarrow \{0, 1\}$ は tolerance class を構成要素と非構成要素に分類する関数である。各要素 $x \in U$ は、$I(x)$ と記される tolerance class を持つ。任意の集合 $X \subseteq U$ について、lower/upper approximations は各々、$\mathcal{L}(\mathcal{R}, X) = \{x \in U \mid \nu(I(x), X) = 1\}$、$\mathcal{U}(\mathcal{R}, X) = \{x \in U \mid \nu(I(x), X) > 0\}$ と定義される。

Tolerance Relation を用いた情報検索モデル

本研究では、equivalence relation の代わりに tolerance relation を用いるラフ集合を応用した情報検索モデルを提案する。研究は 3 つの部分に分けられる。第 1 に tolerance space を定義し、第 2 にラフ集合上の関係を用いたアルゴリズムを定義し、最後にこのアルゴリズムをシステムに適用する。

任意の 2 つの部分集合 X, Y に対して、X, Y の lower/upper approximation が等しいとき rough equality と言い、包含関係にあるとき rough inclusion と言い、重なりあっている (積が空でない) 時 rough overlap という。rough equality と rough inclusion は各々の関係が lower approximation のみの場合、upper approximation のみの場合、及び それら両方の場合があり、rough overlap は lower approximation が重なっている場合及び upper approximation のみが重なっている場合がある。これらを合計して 8 種類の関係が考えられる。

本研究では、データベース中の全キーワードを要素とする集合を全集合とする。文献と質問は共に U の部分集合であり、適合文献の選択は、文献と質問とのラフ集合上の関係を用いて行なう。すなわち、文献と質問を示す集合間の (1) 完全一致 (2) rough equality (3) rough inclusion (質問が文献に) (4) rough inclusion (文献が質問に) (5) rough overlap

の 5 層に対してそれぞれ適合文献を選択し、最終的にそれらを総合して検索文献とする。5 層は更に、関係の種類によって 12 レベルに細分される。

tolerance relation の決定には、文献中のキーワードの共起 (co-occurency) 回数を用いた。2 つの語 t_i および t_j のデータベース中における共起回数を $c(t_i, t_j)$ としたとき、$c(t_i, t_j) \geq \theta$ を満たす t_i, t_j が同じ tolerance class にふくまれることとする。ただしθ は閾値である。

(5) の層において、余りに多数の文献が検索される場合のために、この層に限り、2 次的な順位づけを行なうこととした。この順位づけは単純な集合の積演算によってなされる。

システムへの適用と評価

この手法を用いた実験用検索システムを作成し、評価を行なった。システムは 2 つの部分から成る。まず前半部では、データベースから全ての語の tolerance class と全ての文献の lower/upper approximation を計算する。後半部では質問が与えられ、それに対して適合する文献を探索する。データベースとして、人工知能学会論文誌の過去 10 年間 (1986-1995) の全文献 (802 文献) を用いた。語は、著者によって付与されているキーワードを利用した (1813 語)。

その結果、(1) (2) (3) (4) の各層では、適切な文献が高精度で、しかも適切に順位付けられて検索されることが確認された。また、(5) の層では多数の文献が検索されることも確認された。2 次的な順位づけにより、(5) の層では再現率を保ったままで精度を向上させることが可能であった。

tolerance relation を用いたラフ集合による情報検索モデルは、高精度での検索に用いる際に有用であることが明らかになった。ただし、rough overlap による検索を用いると、時として精度の低い検索が行なわれてしまう場合もあることも分かった。