論 文 の 内 容 の 要 旨

論文題目　　Webにおける情報生産及び利用状況の解析とその応用に関する研究

氏　　名　　佐藤進也

情報は送り手と受け手の間のインタラクションに付随して発生するものであり，元来動的な性質を有するものである．しかしながら，いままでの情報処理は，情報をその生産者や利用者から切り離し，静的なものとして扱ってきた．その大きな理由として次の2つが考えられる．

第 1 の理由として，情報の伝達機構が十分に発達していなかったためその流通は容易でなく，生産者や利用者から独立した状態で管理せざるを得なかったという状況が挙げられる．第 2 の理由として，情報を生産者や利用者から独立したものと考えることにより，検索などの情報処理の問題が簡明化され，解きやすくなるということが挙げられる．情報が生産者や利用者から独立した状態で管理されている状況下では，情報だけに注目することは妥当なアプローチであったと言えよう．実際，文書および検索の問い合わせを語の集合として抽象化し相互の関連性を語の出現に関する統計量などを用いて算出するという手法は，現在も文書の検索や分類などのために広く用いられており，このアプローチは一応の成功を収めた．

しかし，情報がその生産者や利用者とは独立で静的なものとして扱われている状況は，いわゆる情報の電子化と，電子化された情報の流通媒体であるインターネットおよびその上の情報システム，とりわけ World Wide Web（Web）の発展と普及により大きく変化しつつある．たとえば，情報に対してその生産者と利用者がどのように関与しているか（したか）ということも Web の解析によりある程度把握できる．この新たな状況は，情報処理においても新たな可能性をもたらすと考えられる．

その一例として，ハイパーリンクで与えられる Web ページ間の関係を利用した Web 検索を挙げることができる．Web 検索の利用者が検索質問に使う語の数は，平均 1〜2 程度であることが知られている．たとえば，ある利用者が“Java”いう一語で検索したとしよう．この要求に対して，

いままで広く用いられてきた情報検索の手法では“Java”が高頻度で出現する文書を適合性の高いものと判断する．“Java”で検索する利用者のほとんどが Java 言語の情報を必要としていたとしても，この手法によれば，“Java”という語が他ページより多く使われているという理由で，たとえば Java 島について書かれたページが選び出される可能性がある．一方，もし情報の需要と供給に関する知識，たとえば「Java に関しては（Java 言語のオフィシャルサイトである）http://java.sun.com/が参照される頻度が高い」という語と情報の使われ方に関する知識が得られたならば，これは検索結果として文書を選択する際の有力な判断材料になる．そして，実際，ハイパーリンクのアンカーテキスト（リンクで指し示されるページが，リンク元でどのようなことばで参照されているか）を統計的に解析することで，この種の知識が獲得できるのである．

Web が広く普及し社会性を帯びている現在，例に示した手法は，社会における情報の需要に関する動向解析に相当すると考えられる．そして，個人の検索要求を動向解析の結果から推定するということは，個人（あるいは小規模集団）のミクロな情報要求を，社会のマクロな情報要求で近似するということである．個人の活動は社会の趨勢の影響を受けていることを考えると，多くの場合にこのアプローチは適切な結果をもたらすと期待される．この考え方を押し広めると，情報システムは次のようなものとして捉えることができる．すなわち，情報システムとは，人間の社会的な営みとしての情報の生成，流通，利用の特徴をとらえ，それらの効率化を支援する手段と考えることができる．

本研究の目的は，情報システム，なかでも検索システムをこの観点から捉え直し，その新たな可能性を探ることにある．ここで，「検索システム」ということばは，必要とする情報の効率的獲得を支援するシステムという広い意味で用いている．この研究は，従来の情報システムを高度化する方法を探るものとして位置づけることができるが，その意義はそれだけにとどまらない．情報との関わり方において，いま我々を取り巻く状況は大きく変化しつつある．本研究には，その新しい状況への適応方法を示すという重要な意義がある．Web の普及により，多種多様な，大量の情報が自律分散した主体によって動的に生成され、様々な形態で流通・利用されている現在，Web 検索の例で示した通り，従来の手法だけでは必要とする情報の効率的獲得は難しい．この問題に対して，一つの解決の方針を示すことが本研究の目的である．

この問題意識に基づき，本研究では，既存の手法により解決することが難しかった，ことばの意味や情報の価値の把握という課題に焦点をあてる．具体的なアプローチとしては，Web を，情報と，それを生み・利用する者からなるシステムとしてとらえ，その「ふるまい」を観測・解析し，応用することで，意味や価値を把握する方法を考える．本研究の成果は以下の 4 項目にまとめることができる．

(1) 意味体系の動的な生成

Web 検索履歴に記録されている「検索語」と「検索結果中から選択し閲覧した Web ページ」の相互関係を Query Network というネットワーク構造により表現する方法を示す．検索により対応付けられた個々の検索語 t と Web ページ p には，「t の意味は p （の内容）により説明される」「p の内容は t ということばで特徴付けされる」という，ことばと文脈（意味）の相補的関係がある．Query Network では，同一の Web ページ p が t とは異なる検索語 t' で検索されたとき，あるいは，同一の語 t で p とは異なるページ p' が検索されたとき，それぞれ t と t'，p と p' を関

連づけることによりことばと意味を結び付け，体系化する．Query Network は，網羅性や厳密性の点でいわゆるオントロジーのレベルには達していないが，各時点での情報要求に焦点をあて，関連する語彙の収集と整理による「緩い」体系化を動的かつ簡便に行うことができる．この体系を利用して，情報発見・収集を支援できることを示す．

(2) 固有表現に着目した文脈の把握

文書中で語が用いられている文脈を理解するためには，実世界を意識したデータの解釈，具体的には，実世界のエンティティが文書の中にどのように現れ，相互にどういう関係を形成しているかを調べるという方法が考えられる．この考え方を Web における人物の識別に適用し，同姓同名人物の分離を行う．これは，与えられた人名が出現する Web ページを同一人物ごとにグループ分けするタスクで，本手法を用いた場合，高精度で処理できることを示す．この手法では，エンティティ間の相互関係をネットワークにより表現し，その構造の解析により同一性を識別する．ここで注意すべきは，識別の精度がネットワークの構成方法に依存するということである．本論文ではこの事実を指摘し，高い識別精度をもたらすネットワークの構造的特徴を明らかにし，エンティティの相互関係を適切に表現するための基準を示す．

(3) 情報生産者の特徴の時間的側面からの定量化

各 Web サーバへのアクセス状況を，アクセス数とアクセスしているクライアント数の「ゆらぎ」によって把握し，それに基づいて Web サーバの情報生産者としての特性を推定する方法を示す．近年，ネットワーク環境（インターネット，特に Web）の普及にともない，情報は自律分散した主体によって動的に生成・利用されている．このような状況では，情報の属性も動的に変化する．そして，情報の価値も情報の生成や流通のダイナミクスを考慮して判断すべきである．本手法はその要求に応えるものである．

(4) 局所的な情報流通環境の動的な構成

Web サーバに情報発信者のエージェントとしての役割を与え,関連情報を持つ Web サーバを探索させる方法を示す．この方法により，Web サーバのグループが自律的に形成されるが，Web 全体を大域的な情報空間と考えたとき，このグループは局所的な情報空間として捉えることができる．そして，この局所的な情報空間は情報の利用者にとって重要な役割を担っていると考えられる．たとえば，その情報に興味を持っている人の多さ，すなわち「人気」は情報の有用性を測る一つの大域的な尺度であるが，この尺度を単純に適用しただけでは，限られた人々だけが興味を持つ情報の，その人々にとっての価値を測ることができない．この例が示唆するのは，情報の価値は，しばしば，局所的かつ動的に決定されるということである．本論文では，情報の価値判断のみならず，効率的な情報の探索，新しい関連情報の発見を可能とする情報流通環境としての局所的な情報空間の構成方法および利用方法を示す．

これらの成果は，本研究の根本にある「情報と，それを生み・利用する者すべてを一つのシステムとしてとらえ，そのふるまいを観測・解析し，応用することにより，ことばの意味や情報の価値を把握する」という考え方が現実の問題を解くための指針として妥当であることを，複数の

側面からの実証するものである.