平成17年度 情報工学コース卒業研究報告要旨

吉川 研究室	氏　　名	毛 受 崇
卒業研究題目	プレゼンテーション文書から抽出した知識を用いた柔軟な検索に関する研究	

近年, 電子文書の数はますます増加している. 特にインターネット上では, 膨大な数の文書が日々生産, 消費, 蓄積されている. 現在, これらの文書の中から特定の文書を探す場合, ウェブ検索エンジンや情報検索システムを利用することが多い.

従来の検索エンジンや情報検索システムの多くは, 文書の有用性を評価するときに, 内容やリンク構造などは考慮してきたが, 誰が検索を行うのかということは考慮してこなかった. しかし, 同じ検索問合せでも, 利用者が求める検索結果は利用者の知識, 興味, 関心, 嗜好などによって異なるはずである. このような利用者の情報を考慮すれば, 利用者ごとに適切な結果を柔軟に返すことが可能になると考えられる.

そこで本研究では, 利用者のローカルコンピュータに保存されたプレゼンテーション文書集合から知識を抽出し, 抽出した知識を用いて検索問合せを拡張することにより, 検索を利用者に適応させる手法を提案する. 手法の概要図を図 1 に示す. 提案手法では, プレゼンテーション文書において, 異なる概念同士がどのように, そしてどのくらいの強さで関連付けられているかということを知識と見なす. そしてひとつの単語がひとつの概念を表しているものとし, プレゼンテーション文書中に出現する単語同士の共起度を知識として抽出する.

本研究の特徴は, 知識の抽出に先立ち, プレゼンテーション文書をモデル化する点である. 同じ文書内に単語同士が共起する場合でも, 同じページに共起する場合と異なるページに共起する場合とでは, 前者の方が関連が強いと考えられる. また同じページに共起する場合でも, 構造的に近い要素同士で共起する方が関連が強いと考えられる. この「ページ」と「要素」に注目したプレゼンテーション文書のモデル化の例を図 2 に示す.

提案手法の有効性を確かめるために実験を行った. プレゼンテーション文書集合には, データベースの国際会議である VLDB2005 にて発表された文書を使用した. また, 検索対象の文書集合には, IEEE Computer Society の論文を XML 化した文書集合である INEX テストコレクションを使用した.

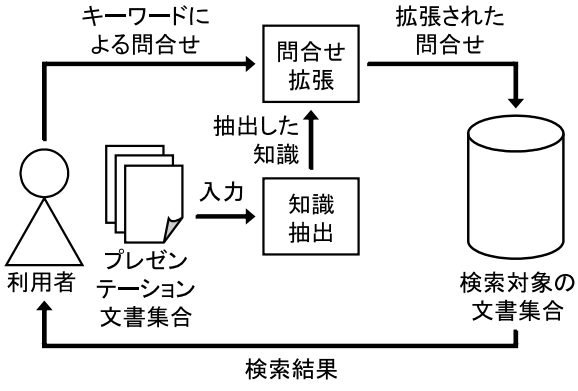

図 1: 提案手法の概要図

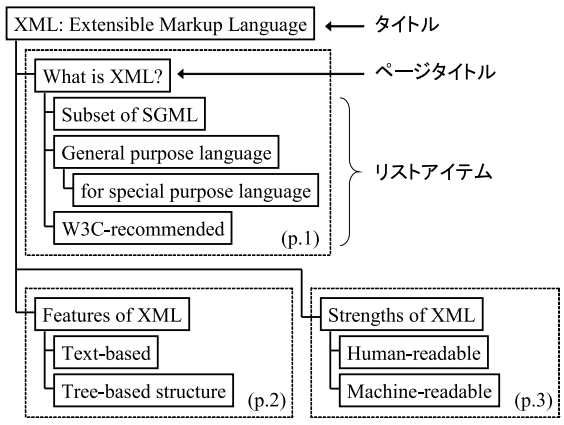

図 2: プレゼンテーション文書のモデル化の例