3G－10

動的シソーラスを用いた連想検索 ～リンク重みの導入～

巌寺俊哲　木本晴夫
ＮＴＴ情報通信網研究所

１．はじめに

近年，コンピュータは著しい技術進歩，普及を遂げている．また，種々の情報のデータベース化も進展し，それらのデータベースは，ネットワークを介して，容易に利用可能になっている．さらに，専門家だけでなく，いままでコンピュータと無縁であった一般ユーザにとってもコンピュータやデータベースが身近に，また，必要不可欠なものになってきており，自らそれを操作する必要が生じている．

従来，データベース検索を行う際には，多くの場合，キーワードをAND/ORで組み合わせた検索式を使用していた．しかし，目的の情報を検索するためには，適切なキーワードを発見し，それらを適切に組み合わせた検索式を作成する必要があり，容易には，目的の情報を検索することができなかった．

現在，我々は，目的の情報を高精度で，容易に，検索することを可能にするために，動的シソーラスを用いた連想検索法の検討を進めている[1,2]．

本稿では，連想検索法の検索精度をさらに向上させるために導入した動的シソーラスのリンク重みとこれを用いたキーワードの関連度の算出法について報告する．また，実験により本方法が適合率の向上に有効であることを示す．

２．動的シソーラスを用いた連想検索

連想検索法は，以下の４点を特徴とする．

（１）サンプル文書の学習
（２）動的シソーラス
（３）連想キーワードの生成
（４）検索結果の順位付け

サンプル文書の学習では，ユーザから提供された検索の目的とする文書の見本（サンプル文書）から検索の目的に関して特徴的なキーワードやキーワードの関係を学習する．さらに，それを動的シソーラス上に反映する．

動的シソーラスは，ノードとリンクとから構成される．ノードは，検索に使用され得る語句を示す．リンクは，その語句間の関係を示す．ノードは，その語句の重要度を示すノード重みを持つ．また，リンクは，語句間の関係の強弱を示す本稿で述べるリンク重みを持つ．

連想キーワードの生成では，ユーザの入力したキーワードから動的シソーラスを使用してユーザの検索目的により適したキーワード（連想キーワード）を生成する．

検索結果の順位付けでは，連想キーワードを使用して検索された検索結果文書を動的シソーラスと照合する．さらに，その動的シソーラスとの適合の程度に応じて順位付けする．

従来の連想検索法では，入力キーワードに対応するノードと関係の強弱にかかわらずリンクで結合してるノードを連想キーワード候補とし，ノード重みを使用して連想キーワードとして生成するか否かを決定していた．この方法では，キーワード間の関係の強弱がまったく考慮されていない．このため，ノード重みを併用せず，リンクのみを連想キーワード生成に使用した場合，再現率は向上するが，適合率は低下する．これは，リンクが，適切なキーワードの生成には有効に機能しているが，不適切なキーワードも生成してしまっているためである．さらに検索精度，特に，適合率を向上させるためには，より適切な関係にあるキーワードのみを生成する必要がある．

３．リンク重みの導入

前述したように，適切な連想キーワードを生成するためには，より適切なリンク，すなわち，関係がより強いリンクを使用する必要があり，その関係の強弱を表現する尺度が必要である．このリンクに付与される関係の強弱を示す値をリンク重みとよぶ．

このリンク重みは，次の条件を満たす必要がある．

（１）リンクは，検索目的に関して，必要な連想キーワードを生成している（再現率向上に寄与）[1]．この機能を損なうことがない．

（２）リンク重みを導入することによって必要なキーワードのみを生成可能にする（適合率の向上）．

上記条件を満たすリンク重みを連想キーワードの生成に使用することで検索精度を向上させることが可能である．

３．１　リンク重みの付与方法

連想検索法では，リンクの生成に共起関係を使用している[1]．この共起関係の頻度をその強弱ととらえ，これを使用してリンク重みを付与する．

リンクの生成と同様にサンプル文書中でのノードに対応するキーワードどうしの共起頻度を使用して算出する．共起頻度とは，ある２つのキーワードが共起したサンプル文書数である．ここで全サンプル文書数がＤＮ，ノードｉとノードｊとに対応するキーワードの共起頻度がＤＮijのとき，ノードｉとｊを結合するリンクのリンク重みＬＷijは，次式により算出される．

$$LWij = \frac{DNij}{DN}$$

上式により算出されたリンク重みは，必ず共起するノード間のリンクでは１に，まったく共起しないノード間のリンクでは，０になる．

３．２　リンク重みを使用した連想キーワードの生成方法

（１）連想キーワード生成手続き

以下に述べる手順で連想キーワードを生成する．

①ユーザの入力キーワードを後述するサンプル文書中での出現確率で点数付けする．

②入力キーワードに対応するノード（生成開始ノード）からあらかじめ指定した距離までリンクで結合しているノードを連想キーワードの候補とする．ここで距離は，生成開始ノードからあるノードまでの経路上にあるリンクの本数である．

③各連想キーワード候補の入力キーワードとの関連度をリンク重みを用いて計算する．関連度は，後述する方法で求める．こ

Associated Information Retrieval using Dynamic-Thesaurus: Incorporation of the Link Weight
Toshiaki IWADERA, Haruo KIMOTO
NTT Information Systems Laboratories

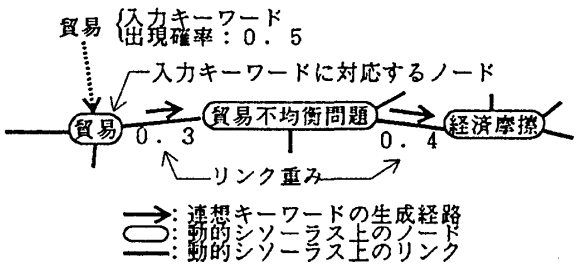

図１　関連度の算出

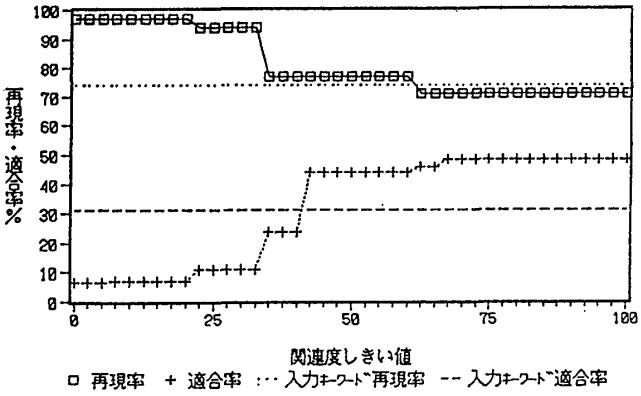

図２　関連度しきい値と再現率・適合率の関係

のとき，複数の入力キーワードから到達し得る候補は，①で求めた点数のより高い入力キーワードとの関連度を採用する．
④あらかじめ指定した値（関連度しきい値）以上の関連度を持つ連想キーワード候補を連想キーワードとして生成する．

ユーザの検索目的に関して，関連度の大きい連想キーワードほど入力キーワードと強い関係を持つ．

（２）サンプル文書中での出現確率

あるキーワードｉのサンプル文書中での出現確率ＯＰｉは，そのキーワードｉが出現したサンプル文書がＤＮｉ，全サンプル文書数がＤＮのとき，次式により算出される．

$$OPi = \frac{DNi}{DN}$$

（３）関連度

入力キーワードｉと連想キーワードｊの関連度ＲＷijは，次式のように，入力キーワードの出現確率ＯＰｉと連想キーワードｊを生成するために通過した動的シソーラス上のリンク（ｉ，…，ｎ）のリンク重みＬＷｉ，…，ＬＷｎの積として算出する．

$$RWij = OP_i \times (LWi \times \cdots \times LWn)$$

たとえば，動的シソーラスの状態が図１のような場合，入力キーワード「貿易」から生成された連想キーワード「経済摩擦」の関連度ＲＷは，次のようになる．

$$RW = 0.5 \times (0.3 \times 0.4) = 0.06$$

４．実験による評価

関連度しきい値の変化とこれに対応する検索結果の再現率，適合率の変化を測定し，リンク重みを用いて算出した関連度の有効性を評価した．

４．１　実験方法

新聞記事１６１件が格納されている文書データベースを対象とし，二人の被験者によって入力されたキーワードから上記方法で生成された連想キーワードを使用して検索実験を行った．学習させるサンプル文書としては，被験者が検索目的に合致していると認定した文書データベース内の文書の半数を使用した．

関連度しきい値は，０～１００範囲で変化させ，その値以上の関連度を持つ連想キーワードのみを検索に使用した．このとき，前述の算出法で求めた関連度は，あらかじめ０～１００の範囲に正規化する．これにより，連想キーワードのもつ関連度の最大値は，１００に，最小値は，０になる．

今回は，リンク重みの有効性を評価するためにノード重みは，使用しない．

４．２　実験結果

実験結果を図２に示す．

実験結果では，リンク重みを使用して算出した関連度のより大きい連想キーワードを使用して検索することによって，再現率は，入力キーワードのみを検索に使用した場合とほぼ同等の値を維持し，かつ，適合率が向上している．

これは，リンクを使用することによって入力キーワードを，ユーザの検索目的に関して，より適切なキーワードへ変換可能であり，そのリンク重みの強いキーワードほど適切な関係にあることを示している．

さらに，上記のことは，共起関係が入力キーワードからユーザの検索目的に対してより適切なキーワードを生成する上で意味があり，その共起関係の強度は，共起頻度から算出し得ることを示している．

５．おわりに

本稿では，より適切なキーワードを生成するために導入した動的シソーラスのノード関係の強弱を示すリンク重みとこれを用いた連想キーワードの関連度の計算方法について報告した．さらに，実験によって関連度の有効性を評価した．その結果，リンク重みが，より適切な連想キーワードの生成に有効であり，検索精度，特に適合率を向上させるのに有効であることが確認できた．さらに，ノード重みと組み合わせて連想キーワードの点数付けに使用することで検索精度をさらに向上させることが期待できる．

［参考文献］

[1]巌寺，木本：「動的シソーラスを用いた連想検索」，情処学会自然言語処理研究会資料90-NL-75, 1990.

[2]H. Kimoto, T. Iwadera : "Construction of a Dynamic Thesaurus and Its Use for Associated Information Retrieval," Proc. of 13rd International Conference on Research & Development in Information Retrieval, ACM SIGIR'90, Brsseles, 1990.