情報システムと社会環境　70－4
（１９９９．３．１６）

Webブラウザを用いた コンコーダンスシステムの開発

井上　康文†、早川　栄一‡、並木　美太郎†

† 東京農工大学　工学部
‡ 拓殖大学　工学部

本稿では、ネットワーク上のテキストに対して、KWIC（Key Word In Context）による参照や検索を目的としたシステム開発について述べる。文学書や新聞がネットワーク上に公開され、Webブラウザを使ってテキストを参照することが多くなったが、フルテキストの文書に対して、検索の機能だけではネットワーク上の必要な情報の参照に時間がかかっていた。そこで本システムは、ネットワーク上のテキストに対してコンコーダンスを作成し、そのコンコーダンスデータを使うことでネットワーク上のテキストの参照を補助することができた。またシステムをJava言語で記述したことで、Webブラウザ上でコンコーダンスの作成から参照までを行うことができた。

Development of a Concordance System on Web Browser

INOUE Yasufumi †, HAYAKAWA Eiichi ‡, NAMIKI Mitarou †

† Department of Computer Science, Tokyo University of Agriculture and Technology
‡ Department of Computer Science, Takushoku University

This paper describes the development of a system with concordance that serves referring, and retrieving documents on the Internet. Documents that are public on the network, such as literature and newspapers, are frequently referred with Web browser, but retrieving function like search engines cannot present quick reference for necessary information from large quantities of documents on the network. We developed a system that makes concordance from documents on the network and uses it as the support of information retrieval. The system utilizes referring contents quickly through concordances. All the system is written in Java, so that it can be executed on any Web browsers.

1．はじめに

ネットワークが急速に発達し、本来紙の上での情報であった文学書や論文、新聞やニュースなどの文書が電子化され、ネットワーク上に公開されるようになった。しかし、テキストは必要としている個所を探すのに手間がかかるため、ネットワーク上の文書から必要な部分を探す場合、既存のサーチエンジン[6][7][8]や Web ブラウザ[9][10]等の検索の機能を使うだけでは、候補が多く、時間がかかっていた。

文学や歴史学の研究では、コンコーダンスを参照したい内容の記述個所を探すために利用して、研究に役立てている[1][4]。コンコーダンスとは、語句の出現する位置と KWIC (Key Word In Context) が記載されている索引である。コンコーダンスを使うことで、語句の前後の文脈からその語句に対する情報を読み取ることができ、文書中の必要な部分を探しやすくなる。しかしコンコーダンスは、語句の全文検索や文脈の切り出しの作成に非常に時間がかかるという問題点がある。

そこで我々は、コンコーダンスの作成を個人でできるよう、語句の検索や文脈の切出しを自動的に行い、さらにその作成したコンコーダンスを用いることで、ネットワーク上にあるテキストの参照や検索の補助を行うシステムを考案した。本稿では、ネットワークに対応したコンコーダンスシステムの設計と実現について述べる。

2．問題分析

2．1 紙面上でのコンコーダンスの問題点

紙面上でのコンコーダンスの作成や参照の問題点として、次のことが挙げられる。

（1）語句を引くなどの作業に時間かかる

紙面上の文書の索引では、文書中での語句の出現位置を表す方法として、ページ番号や行数、段落などが使われている。語句の出現位置を引くには、膨大な索引の中から必要な項目、ページや行などを探すという作業が必要で、面倒である。

（2）コンコーダンスの作成に時間がかかる

ある語句から文書を参照しようとした場合、コンコーダンスの KWIC を使うことで、文書中でのその語句についての情報を得ることができる。しかし、個人でコンコーダンスを作成するには、KWIC の作成や語句の全文検索など時間のかかる作業である。

2．2 ネットワーク上のテキストを参照する場合の問題点

ネットワーク上の文書を参照する方法として、Web ブラウザによる参照方法が挙げられるが、その場合での問題点は次のとおりである。

（1）キーワードから直接参照したい個所を見つけにくい

キーワードからネットワーク上の必要な情報を探す場合、既存のサーチエンジンなどが使われるが、キーワードの出現数や文書の始めの部分の切出しだけでは、文書の内容が必要な情報かどうかわからない。また、文書中のキーワードが出現する位置を、直接参照することができない。

（2）ネットワーク上の文書に対して、新しい情報の書込みができない

ネットワーク上の文書は、読出しのみのものが多く、書込みができない。書込みができないことで、一度参照した文書に対して個人の情報や参照した結果を残すことができず、個人向けに再利用を考えた整理が難しい。

3．システムの設計目標と方針

3．1 設計目標

本システムの設計目標を次に述べる。

（1）コンコーダンスを用いることでテキストの内容に対する参照の補助をする

Web ブラウザなどに付いてくる単純な文字列検索だけでは、テキスト内の必要としている個所を簡単に見つけることができず、効率的に参照することは難しい。コンコーダンスを用いることで、キーワードから直接必要としている情報を探し、テキスト参照の補助をすることができる。

（2）ネットワークを介したコンコーダンスの作成とテキストの参照を可能にする

本システムを使うことで、ネットワーク上のテキストに対してコンコーダンスを作成でき、その作成されたコンコーダンスデータを使い、ネットワークを介したテキストの参照を補助する環境を提供する。

（3）個人レベルでコンコーダンスを作成できる

システムを計算機上で実現することで、検索やソートの部分を計算機で行い、個人でコンコーダンスを作成するための補助ができる。その結果、人の手で行うよりも早くコンコーダンスを作成できる。

（4）コンコーダンスを共有できる

一つの文書に出現するすべてのキーワードに対してコンコーダンスを作成するとなると、時間や労力が必要になる。しかし、作成したコンコーダンスを共有することで、他の人が作成した既存のコンコーダンスデータを使った参照や、新しい情報を追加という形でコンコーダンスのデータを作ることができる。

（5）環境に独立する

ネットワークを使ったシステムを作成する場合、そのシステムがどの環境に対応しているのか問題となる。しかし、ネットワーク上にはいろいろな環境が存在し、Web ブラウザを使ったネットワークの参照と同じように、いつどこで行われるかわからない。そこで本システムは、ユーザの環境に左右されないで使うことができる必要性がある。

3．2 設計方針

（1）ネットワーク上のフルテキストの文書に対してのコンコーダンスの作成

ネットワーク上のファイルは、書込みができない。そこでネットワーク上のテキストには、コンコーダンスの追加や書込みは行わず、作成したデータと分離したデータとして扱うことで、ネットワーク上のテキストに対してコンコーダンスを作成できるようにする。

（2）サーバ・クライアント型のシステム

本システムで作成されるすべてのデータを、ユーザに対して共通のサーバに保存し、クライアントからそのデータにアクセスすることで、同じコンコーダンスデータを複数のユーザで使うことができる。

（3）Java 言語によるシステムの記述

システムを Java 言語で作成し、コンコーダンスの作成やコンコーダンスを使った参照の部分をアプレットとして作成することで、システムを起動するという手間を考えることなく、Web ブラウザ上でコンコーダンスの作成や参照でき、異なった環境でもシステムを使うことができる。

4．システムの設計

4．1 システムの特徴

本システムの特徴を次に挙げる。

（1）KWIC の自動生成

ユーザからコンコーダンスを作成するキーワードが入力されると、システムは文書中

を全文検索し、自動的にKWICを生成する。紙面上のでのコンコーダンスでは、このKWICの作成に時間が必要だったが、本システムでは自動的に行うことで、コンコーダンスの作成の時間を短縮することができる。

（2）文書からコンコーダンスを引くことができる

紙面上のコンコーダンスでは、コンコーダンスから文書中のキーワードを引くことができるが、本システムでは文書中のキーワードからコンコーダンスへの逆引きもできる。本機能により、文書からすぐにコンコーダンスを参照できる。

（3）複数人でコンコーダンスを作成できる

コンコーダンスを共有したことで、共同作業としてコンコーダンスを作成し、一つの文献に対して複数人で研究できる。

（4）キーワード検索

ネットワーク上のテキストに対して、キーワードだけの情報から必要な個所を探す場合、検索エンジンなどが使われるが、検索結果から文書の内容が読み取れないため、必要としている文書を見つけるのが難しい。そこで本システムは、コンコーダンスが作成されたキーワードに対する検索機能を持ち、検索結果にKWICを表示することで、従来の検索エンジンよりも、文書中でのキーワードの意味や使い方を知ることができ、必要としている文書が調べやすい。

4．2 全体構成

本システムは、ユーザからの入力やユーザへのデータの表示を行うコンコーダンス部と、ネットワークとコンコーダンス部のデータのやり取りを仲介するファイル管理部とに分けることができる。全体構成を図1に示す。

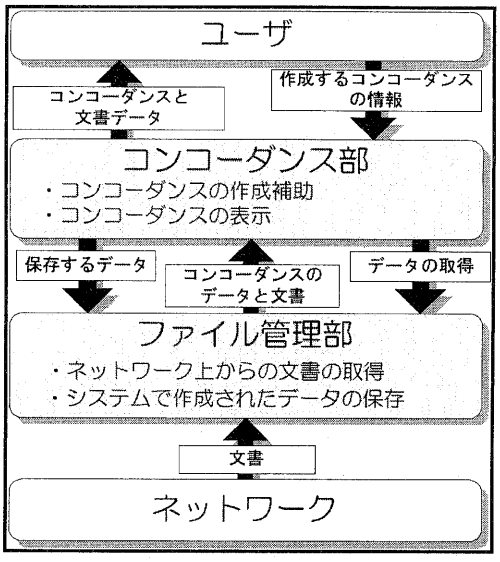

図1．システムの全体構成とシステムで作成されるデータの流れ

（1）コンコーダンス部

コンコーダンスのデータへのリンク情報を持った文書（以下「文書データ」）とコンコーダンスのデータ（以下「コンコーダンスデータ」）の表示、ならびにコンコーダンスを作成する。アプレットとしてWebブラウザ上で実行される。

（2）ファイル管理部

コンコーダンス部から要求のあったファイルをネットワーク上から取得し、コンコーダンス部へと渡す。また、コンコーダンス部で作成されたコンコーダンスデータや文書データを保存する。

4．3 サーバ・クライアント型システム

このシステムは、作成されたコンコーダンスを共有するため、クライアント側で作成されたコンコーダンスをネットワークなどを通して、サーバへ送られ保存される。クライアントはWebブラウザ上でアプレットとして動いているコンコーダンス部、サーバはファイ

ル管理部となる。

サーバとクライアントの関係を図２に示す。

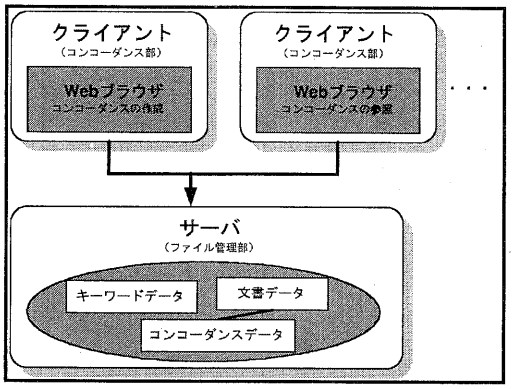

図２．サーバ・クライアント関係図

４．４ システムで管理されるデータ

本システムで作成され、管理されるデータは次の３種類ある。

（１）コンコーダンスデータ

キーワードの KWIC と文書中の全出現位置の情報、コメントや関連付けなどのデータ。

（２）文書データ

ネットワーク上の文書にコンコーダンスデータへのリンク情報を持った、新しい文書。

（３）キーワードデータ

作成されたコンコーダンスのキーワードをまとめたものである。キーワードを検索するときに使う。

これらのデータの使い方は次節で説明する。

４．５ システムで作成されるデータの流れ

本システムは、ネットワーク上のテキストに対して、コンコーダンスの作成から参照までを支援する。その流れを図１に示し、次に各部のデータの流れを説明する。

（１）コンコーダンスの作成

コンコーダンスを作成には、まず対象となるファイルの URL の指定が必要である。ユーザから指定のあった URL をコンコーダンス部が受け取り、ファイル管理部へと渡される。ファイル管理部は、ネットワーク上から目的のファイルを取得し、その内容をコンコーダンス部へと渡す。コンコーダンス部は、そのファイルを表示し、ユーザから作成するコンコーダンスのキーワードを受け取り、自動的にコンコーダンスを作成する。作成されたコンコーダンスは、ファイル管理部へと渡され、保存される。

（２）コンコーダンスを使った参照

文書の参照中、コンコーダンスを表示しようとすると、コンコーダンス部はファイル管理部からコンコーダンスデータを受け取り、ユーザへと表示する。ユーザから、コンコーダンスの変更または追加の情報があると、更新としてファイル管理部へ新しいコンコーダンスデータと文書データを送り、保存される。

（３）作成されたコンコーダンスのキーワードの検索

サーチエンジンのように、調べたい情報のキーワードをコンコーダンス部へ入力すると、ファイル管理部はキーワードファイルから、サーバに保存されているすべてのキーワードのコンコーダンスを探し出し、コンコーダンス部へと送り、表示する。

４．６ コンコーダンスの設計

本システムで作成するコンコーダンスの項目の例として、「アプレット」というキーワードでコンコーダンスを作成した場合の例を図３に示し、説明する。

（１）KWIC

KWIC は、「アプレット」というキーワードが出現する文脈中のすべての位置を切り出したものである。キーワードの文脈上の位置やユーザが必要としている情報によって、表示

する長さを変えることができる。

（2）コメント

「アプレット」というキーワードに関して、KWIC だけではわからない説明やユーザのメモなど、文書を参照してわかったことを書き込むことで、個人での情報の整理とコンコーダンスの参照に役立てる。

（3）関連付け

キーワード「アプレット」とネットワーク上のデータとを関連付ける。ネットワーク上にはテキストだけではなく、音や画像、動画などのマルチメディアデータがある。テキストだけでは表現できないことを、キーワードとファイルという関係を作ることで、参照の幅を広げることができる。

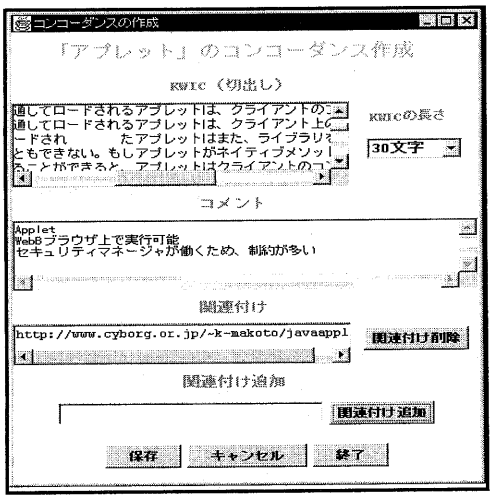

図３．コンコーダンスの項目の例

コンコーダンスのデータは図４のようにHTML で記述され、<PARAM>タグを使うことでアプレットにデータとして渡される。

KWIC をそのまま記述することで、アプレットの起動時に切り出し部分の検索やファイルの読込みをすることなく、表示ができる。

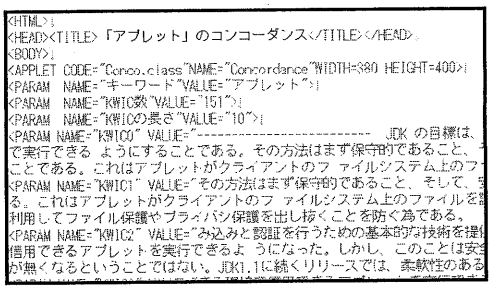

図４．コンコーダンスデータの内容

４．７ 文書データの設計

本システムで「アプレット」というキーワードで文書データを作成し、Web ブラウザで表示した例を、図５に示す。

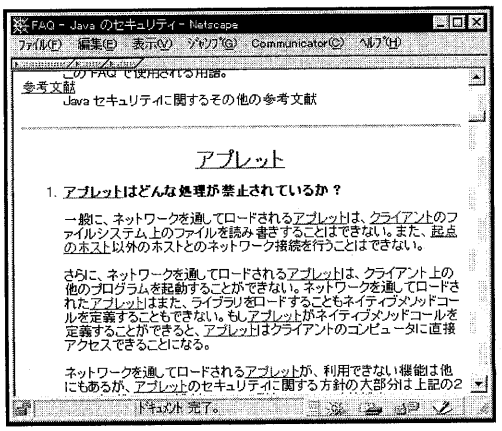

図５．文書データを Web ブラウザで表示した例

紙面上のコンコーダンスでは、キーワードとそのキーワードが出現する文書中での対応は、ページ数や行数で表され、キーワードと文書を対応しながら参照するには、ページや行の情報を手がかりに探す必要があった。また、計算機ではクライアントの環境によって、文章の表示方法が異なり、その記述は使えない。

本システムでは、文書データを HTML で記述し、<A>タグを使うことで、Web ブラウザ

上では文書からキーワードへの逆引きのリンクとして、直接コンコーダンスを参照することができる。コンコーダンスを参照するために、アルファベット順や五十音順に規則的に並べられたキーワードを探す手間と時間をなくし、複数のキーワードのコンコーダンスを同時に表示することができるため、紙面上のコンコーダンスよりも簡単に参照することができる。

文書データとコンコーダンスデータの関係を図6に示す。

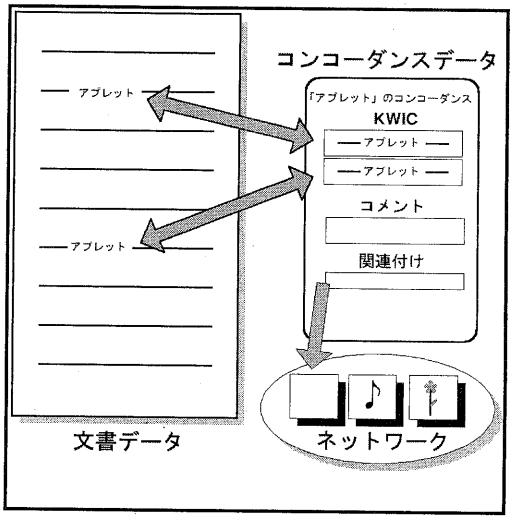

図6．コンコーダンスデータと文書データの関係

5．実現と評価

5．1 使用環境と使用言語

システムを表1の環境で実現した。

表1．開発環境

CPU	Pentium 200MHz
Memory	64Mbyte
言語	Java 言語
開発環境	JDK（Java Development Kit）1.1.7A
ソース行数	約3000行

5．2 システム評価

本システムを用いて、ネットワーク上の文書に対して、コンコーダンスを作成した。対象となった文書の内容は表2のとおりである。

表2．コンコーダンス作成の評価文書

テキスト数	14
最小テキストサイズ	約10Kbyte
最大テキストサイズ	約210Kbyte
総テキストサイズ	約600Kbyte
コンコーダンス作成数	98
総コンコーダンスデータサイズ	約80Kbyte
作成時間	約2時間
作成環境	Netscape Navigator 4.04 + JDK 1.1 support

コンコーダンスの作成時間は約2時間である。ほとんどは著者がコンコーダンスを作成するキーワードを探す時間やコメントを入力する時間であった。紙面上では、KWICの作成や文書の全文検索で、明らかにこれ以上の時間がかかる。

またコメントの項目に書込みや関連付けをすることで、ネットワーク上の文書に個人の情報を付けることができ、キーワードの理解と参照の幅を広げる手助けとなった。

検索速度の結果を表3次に示す。

表3．検索速度結果

ファイルサイズ	1.5Mbyte
KWIC作成時間	約15秒

KWICを作成する全文検索を15秒で行えた。本システムは、Internet Explorer 3.02 +

Microsoft Virtual Machine 3.1 の環境でも動くことを確認した。

6．おわりに

本稿では、Web ブラウザ上でコンコーダンスを作成、参照可能なシステムの設計と実現について述べた。本システムによって、ネットワーク上のテキストに対する参照と検索の補助を行うことができた。

今後の課題として、ネットワーク上ファイルとサーバに保存しているファイルとの動的な対応、サーチエンジンを使った自動コンコーダンス作成によるデータベースの拡張などが挙げられる。

参考文献

[1]坂口基彦他:文科系研究支援のためのコンコーダンスを用いた文書研究システムの設計と実現, 第 54 回情報処理学会全国大会, 1997

[2]福原知宏他:ディジタル図書館における情報作成支援環境, 情報処理学会研究会報告書 情報システム, 63-2, 1997

[3]遠藤教昭他:汎用的な sybperl を用いた図書館データベースと WWW の連携, 情報処理学会研究会報告書 情報システム, 64-3, 1997

[4]小林康夫他:知の技法, 東京大学出版会, 1994

[5]日本サン・マイクロシステムズ:Java Technology, URL: http://www.sun.co.jp/java/, 1997

[6]株式会社エヌ・ティ・ティ・アド他:goo, URL: http://www.goo.ne.jp/

[7]Infoseek Corporation:Infoseek Japan, URL: http://japan.infoseek.com/

[8]Yahoo Japan Corporation:Yahoo! JAPAN, URL: http://www.yahoo.co.jp/

[9]Netscape Communications Corporation : Netscape Navigator, URL:http://help.netscape.com/

[10]Microsoft Corporation : Internet Explorer, http://www.asia.microsoft.com/windows/ie_intl/ja/