卒業論文 1998 年度 (平成 10 年度)

コミュニティの知識を反映した全文検索システムの設計と実装

慶應義塾大学環境情報学部
藤田 桂
katsura@sfc.wide.ad.jp

指導教官
村井 純
徳田 英幸
楠本 博之
中村 修

平成 11 年 2 月 3 日

概要

インターネットの普及により, 国内外のあらゆる情報がWWW上に発信されることが一般的となった. このような中で, インターネット上の膨大な情報から自分の欲しい情報を探してくれる検索サービスが生まれ, 今や多くのサイトでサービスが行なわれている.

現在のインターネット上の全文検索システムの多くは, 対象文書中に含まれる単語の出現頻度のみを利用して検索を行なっている. この方法では検索の再現率は高いものの, 適合率が低いため得られた意味内容に差異のない検索結果の中から本当に自分の欲しい情報を取り出すのが困難である. このときに対象を絞りこんでいく手法として, 論理演算などが主流であるが, 論理演算に馴染みのないユーザにとっての利用は困難である. なぜならばインターネット上に提供された情報とユーザの検索要求とを仲介させる機能が, 既存の全文検索システムでは欠陥しているからである.

本研究ではこの欠陥を補う全文検索システム「Community Knowledge」の実装・設計・評価を行なった. 膨大なインターネット上の情報量に対応するため, ネットワーク上のコミュニティの知識を共有・反映させ, 検索アルゴリズムを出現頻度ではなく文書構造に着目し, 情報の意味内容の提示や意味内容の差異に基づく適合性の高い検索結果をユーザに返す全文検索システムを提案した. 全文検索システムの仲介機能として備えることによって, ユーザは自分の欲しい情報を手に入れることが可能となる.

慶應義塾大学 環境情報学部
藤田 桂
katsura@sfc.wide.ad.jp

Abstract

The spread of the Internet makes it general to open every information at home and abroad to the public on the World Wide Web.Then a search service to find out what we really want from extremely wide information on the internet,was suggested and get often used.

Many of present algorithms of search engines depend on only the frequency of appearances of a word in a target document. This algorithm results in a good recall,but it is difficult to choose what a user really wants without contents of meaning. Though logical operation is often used to narrow targets down, some of users aren't familiar with it. That is, exiting search engines don't have a flex recognition of users' demands for searching.

In this research has designed, implementable and assessed the algorism.Faced to extremely wide information, the search engines use the knowledge of network community and reflects it, independent of the frequency, show the suitable results with the cluster and the difference of meaning. After a search engine gets a flex recognition, a user can obtain information human really wants.

Faculty of Environmental Information, KEIO University
Katsura Fujita
katsura@sfc.wide.ad.jp

目 次

1 序論 **1**

1.1 はじめに 1

1.2 研究目的 1

1.3 本論文の概要 2

2 既存の WWW 上の全文検索システムの手法および問題点 **3**

2.1 本章の概要 3

2.2 WWW 上の全文検索システムの構成 3

2.2.1 基本的な構造 3

2.2.2 ロボットとは 4

2.2.3 文書フィルタ 4

2.2.4 インデックスの作成 5

2.2.5 形態素解析 5

2.2.6 検索手法 6

2.2.7 検索結果 6

2.3 問題点 6

2.3.1 ノイズの多い検索結果 6

2.3.2 キーワードマッチによる検索 7

3 全文検索システム 「Community Knowledge」] の設計 **9**

3.1 本章の概要 9

3.2 問題点からの検討 9

3.3 ノイズの多い検索結果への解決 9

3.3.1 概念の形成 9

3.4 完全照合手法への解決 10

3.4.1 文書の構造に着目した検索手法 10

3.4.2 検索結果表示の改善 10

3.5 全文検索システム 「Community Knowledge」の設計 11

3.6 全文検索システム 「Community Knowledge」の処理の流れ 11

3.6.1 概念の形成 11

3.6.2 シソーラスの作成 12

3.6.3 単語間の類似性を抽出 12
3.6.4 回収した文書の概念によるフィルタリング 14
3.6.5 シソーラス展開による文書のグループ分け 14
3.6.6 検索結果におけるシソーラスの提示と意味内容による差異づけ . . . 15

4 システムの実装 **17**
4.1 本章の概要 17
4.2 環境 17
4.2.1 形態素解析システム「すもも」について 17
4.3 システムの実装 18
4.3.1 ネットワークコミュニティの情報活動の記録を回収 18
4.3.2 コミュニティ辞書の作成 18
4.3.3 単語関連辞書の作成 21
4.3.4 WWW 上の文書を回収 (ロボット部) 22
4.3.5 回収した文書の概念によるフィルタリング 22
4.3.6 文書のグループ分け 22
4.3.7 検索処理 23
4.3.8 検索結果の例 23

5 評価 **25**
5.1 本章の概要 25
5.2 適合率の評価 25
5.2.1 適合率と再現率 25
5.2.2 評価 27
5.3 順位付けの評価 27
5.4 グループ提示の評価 29
5.5 考察 29
5.6 分析 30

6 結論 **32**
6.1 まとめ 32
6.2 今後の課題 32
6.3 おわりに 33

謝辞 **34**

参考文献 **34**

図目次

2.1 全文検索システムの構成 4
2.2 形態素解析の例 5

3.1 単語の類似性 14
3.2 文書群のグループ分け 15
3.3 グループの提示 16

4.1 コミュニティ辞書の作成 19
4.2 単語の抽出 20
4.3 検索結果 24

5.1 再現率 25
5.2 適合率 26
5.3 再現率と適合率の関係 26

表目次

4.1 コミュニティ辞書 21
4.2 単語関連辞書 22

5.1 適合率の評価 27
5.2 出現頻度のみのスコアリング 28
5.3 関連キーワードによるスコアリング 28
5.4 グループの提示 29
5.5 分析用のコミュニティ辞書 30
5.6 分析結果 31

第 1 章

序論

1.1 はじめに

近年, ネットワークを用いて関心を共有するコミュニティの活動を支援する試みが始まっている. このようなコミュニティの特徴として, 静的 ·定型的な情報に加え, メンバが情報活動 (調べる, 質問する, 人を知る, 議論する) によって作り出す動的な情報が生成されることが挙げられる. この動的な情報は専門性が高く, 有益であり, 情報を発見 ·理解 ·整理するための知識の源となる. コミュニティではこの知識を共有 ·蓄積 ·再利用することが重要となるが, 既存の全文検索システムをはじめとする全文検索システムでは, このようなコミュニティの知識を反映するためのしくみを持っていないため, ユーザは情報の発見, 理解, 整理にコストを費してしまう. よってネットワーク上のコミュニティの知識を反映した全文検索システムが必須となる.

1.2 研究目的

本研究は, インターネットをはじめとする, コンピュータネットワーク上において共通の興味を持つ人々の集団に焦点をあてる. ここではこのような集団を「ネットワークコミュニティ」と呼ぶ.

本研究では, 検索プロセスにおいて対象文書中に含まれる単語の出現頻度のみを考えず, 含まれる単語と関連した単語の関係の割合で適合率を高める. また検索結果として返すとき, 対象文書が内容を示す文書構造をを提示することによって, ユーザの文書の選択指標とする. 単語の関連性, 文書の構造の提示を行なう際に, 本研究はネットワーク上のコミュニティが動的に生成する知識を利用した. ネットワーク上のコミュニティ知識を反映させた全文検索システムを実現することによって, ユーザは情報の発見, 理解, 整理をより容易に行なえる.

1.3 本論文の概要

本論文では, 第2章で既存のWWW上の全文検索システムの手法, 問題点に関するレビューを行なう. 第3章ではネットワークコミュニティの知識を反映させた全文検索システムの設計について述べる. 第4章では, 本研究での全文検索システム『Community Knowledge』の実装に関して解説し, さらに. 第5章では, 本研究での全文検索システム『Community Knowledge』の評価を行なう. その結果をもとに, 第6章では結論と今後の課題について述べる.

第 2 章

既存の WWW 上の全文検索システムの手法および問題点

2.1 本章の概要

WWW 上の全文検索システムとは,Web ブラウザの一種である「ロボット」と呼ばれるエージェントプログラムを利用して, インターネット上の WWW ページを定期的に収集し, それらに対するキーワード付けを自動的に行なって, 検索サービスを提供する仕組みである.

本章では, まずこの全文検索システムを構築するための構成を順に説明し, 次に既存の全文検索システムにおける問題点を述べる.

2.2 WWW 上の全文検索システムの構成

2.2.1 基本的な構造

「WWW 上の全文検索システム」は, 以下の要素により構成される.

- リソース収集部 (ロボット)
- 文書フィルタ部
- 取得したデータをまとめて処理するインデクサ
- 検索クライアント
- 検索エンジン部

WWW 上の全文検索システムには, 大きく分けると 2 つの機能がある.1 つは, ロボットにより, 回収された文書に対して, インデックス作成 (インデクシング) を担当する「インデクサ」と呼ばれる部分, もう一つは, インデックス生成で作成されたデータ (インデックス) を利用して, 実際に検索を行なう「サーチエンジン」と呼ばれる部分である.

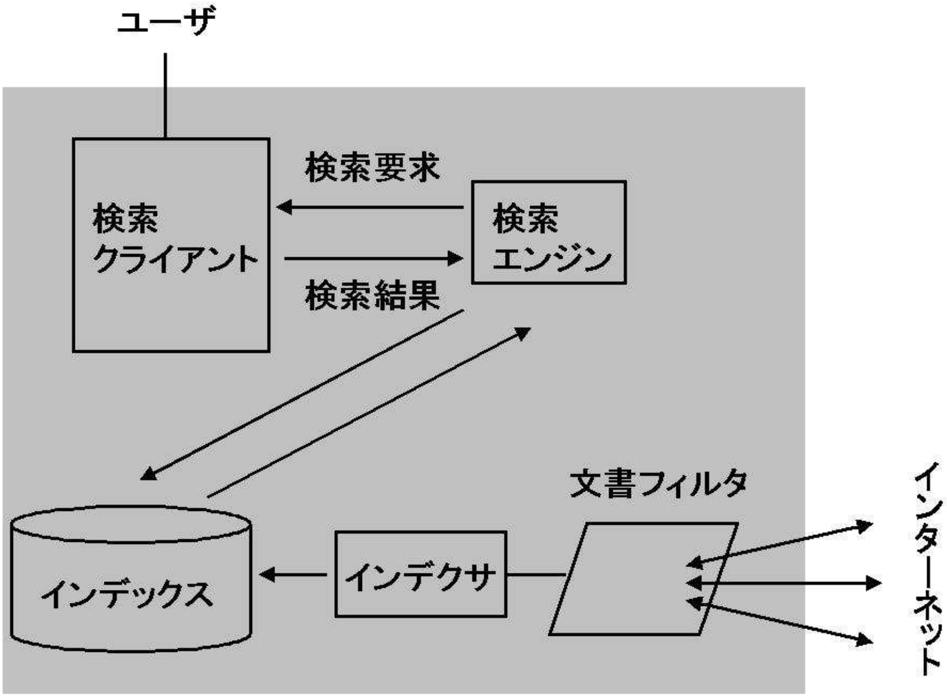

図 2.1: 全文検索システムの構成

2.2.2 ロボットとは

WWW 上の全文検索システムにおける「ロボット」とは,「Web 上のオブジェクトを自動的に取得するプログラム」である.

- AltaVista の scooter[7]
- Lycos の spider[10]

なども, すべてロボットである. 場合によっては「エージェント」と呼ばれることもある.

ロボットのやっていることは,Web ブラウザ上のリンクをマウスでクリックするという人間が行なう動作を, 自動的にやってくれるものである. つまり, ロボットは Netscape Navigator のような Web ブラウザと同じく,Web クライアントの一種にすぎない.

2.2.3 文書フィルタ

WWW 上の全文検索システムを構築するには, 対象データをある一定のディレクトリ以下に集めておく必要がある. 収集した文書ファイルを, インデクサが理解できる形式 (シンプルテキスト) に変換する部分のことを文書フィルタ部という.

2.2.4 インデックスの作成

インデックス作成とは, 実際の検索作業を始める前にあらかじめインデックスを作成しておく作業である. インデックスとは「索引」のことであり, 全文検索システムの場合, インデックス作成で作られるインデックスは, どの単語 (キーワード) がどの文書に含まれているかを示すものである. 全文検索システムでは通常, 文章に含まれているすべての単語についてインデックスを作成する.

2.2.5 形態素解析

全文検索システムでは通常, 単語単位に検索を行なう. そのため, インデックスには単語単位で情報を格納する. したがって, インデックスを作成するには, 文を単語ごとに分ける処理が必要になる. 通常この処理は, インデクサの内部であるアルゴリズムに基づいて行なわれる. 例をあげると, 「藤田君は日本語が下手です」という文章は,

藤田 君 は 日本語 が 下手です

のように単語ごとにスペースで分けられる. これを自然言語学の用語で言うと, 「形態素解析」という.

```
% echo "藤田君は日本語が下手です" | sumomo -p
        "文頭"
藤田 "固有名詞"
君 "接尾語"
は "助詞"
日本語 "普通名詞"
が "助詞"
下手 "普通名詞"
です "助動詞"
```

図 2.2: 形態素解析の例

日本語は欧米語のように単語と単語で空白で区切られていないので形態素解析のソフトウェアが必要となる. 日本語形態素解析システムとして実際に用いられているもののうち, 比較的よく知られているのは,

- JUMAN(京都大学)[15]
- ChaSen(奈良先端技術大学)[19]
- JENOMI(図書館情報大学)[16]

- BreakFast(富士通)[17]
- SuperMorpho-J(オムロンアルファテック)[18]
- すもも(日本電信電話株式会社)[20]

などである.

2.2.6 検索手法

全文検索システムを利用して文章を探すためにはユーザが検索したい内容を表現している単語を入力して使うのが一般的である. 単一単語で検索する場合, あまりにも多くの情報が検索結果として返る場合が多い. その中には関係ない文章までが大量に検索結果として返ることもある. その場合, 複数単語でAND検索,OR検索,NOT検索を行ない, 検索を効率化する.

2.2.7 検索結果

一般的に全文検索システムでは, キーワードを含む文章に対して, それぞれの文章がどの程度検索条件に適合しているかを示す指標として「スコア」と呼ばれる点数付けを行ない, そのスコアの順に並び替えたものを検索結果として出力する. これをスコアリングと呼んでいる. スコアリングは一般的に検索語の頻度を利用して行なう. つまり, キーワードとなる文字列が複数回出現している文書には高い点数が与えられる.

例えば,「FreeBSD」というキーワードで検索した場合,「FreeBSD」という語を多く含む文書ファイルが求められる文書により近いと判断され, 高得点が付けられて上位に位置する. 逆に「FreeBSD」が一度しか表れないような文書には低い点数しか与えられず, 下位に位置する. 全文検索システムはその結果を, その順序でユーザに対して提示するようになっている.

2.3 問題点

2.3.1 ノイズの多い検索結果

全文検索システムを利用するに当たって, 多く見られる問題が, 多くのノイズが混ざってしまう, つまりユーザの意図した文書とは全く関係ない文書が検索結果として返ることである.

ノイズの多い検索結果の対策として挙げられるのが,「Mondou」[6] や「ODIN」[12],「Excite」[13] などで見られる suggestion 機能 (関連語提示機能) である.suggestion の表示により, ある程度ユーザは指標に沿った文書を選択することが可能となる.suggestion 機能のアプローチとして, ユーザが過去に入力したキーワードとその検索結果の履歴を利用して, より絞りこんだ検索結果を返すと言う手法である.

1. ユーザがどのようなキーワードでどの文書を選択したかを記録していく
2. 1. の結果を集計して, 各文書がどういうキーワードで選択されたかの表を作っておく
3. 検索時に, 検索結果となる文書の多くに対して選択するために使われたキーワードを,suggestion として表示する

しかし,suggestion 機能にも大きな問題点が挙げられる.

- ユーザは最初の段階では, ノイズの多い検索結果から文書を探すことになる

  suggestion 機能が実装された全文検索システムは,suggestion を表示するため, ユーザの検索履歴を蓄積しなければならない. しかし蓄積段階でユーザは suggestion 機能がない状態で検索結果から文書を探さなければならないので, ユーザの意図した文書とは違う文書を見なければならない可能性も出てくる. このような段階を踏むと, キーワードと文書の相関性が必ずしも正しいとは言えないものとなってしまう恐れが出てくる.

- 各ユーザの要求によってキーワードと文書の関連性が変わってしまう.

  例えば, 検索キーワードで「ボール」と入れた場合, それが「サッカー」のボールなのか,「野球」のボールなのかもしくは調理器具の「ボール」なのかがわからない. ユーザの要求する目的, 趣向によってキーワードと文書の関連性が全く異なるものになるという問題点が挙げられる.

膨大な WWW 上の文書に対して, 趣向, 目的の異なったユーザそれぞれの要求に応じたキーワードと文書の関連性を求めるのは難しいことであることがわかる.

2.3.2　キーワードマッチによる検索

全文検索システムによるキーワードマッチによる手法, すなわち完全照合手法はユーザの欲しい情報が検索式と同じであり, また, 求める情報を提供する文書の表現やテキストとも同じであることを仮定している. この手法では, 検索質問が検索式の中で表現されていると同じように, 文書表現に含まれていることが必要である. 現在ある全文検索システムは, 完全照合手法に基づく実装であり, ブール検索や文字列検索がごく普通に実現されている.

しかしユーザにとって自分の欲しい情報を概念化し, それをうまく言語化, すなわちキーワードで複数羅列して表現することは難しい. それはユーザの実際の心の中で表現される要求と実際の検索式との間には, 概念的な距離が存在するということである. またそれが提供された文書とうまく適合させなければいけないという課題もある. すなわち情報を探すユーザと提供する情報を仲介させる機能, 提供された文書へ導く課程が欠陥しているという問題点が見られる.

このような既存の全文検索システムで実装している完全照合手法の問題点を以下に示す.

- ユーザの求める情報に適合する文書であっても, その表現が検索式と部分的にしか一致しない場合は, 適合したものとみなされない.

  例えば, ユーザがオリンピックのサッカー情報が欲しい場合, 「オリンピック AND サッカー」というように AND 検索を行なった場合, 「Olympic のサッカー情報」というような 「オリンピック」を 「Olympic」と書かれた文書は除外されてしまう. 提供する文書表現をそのままキーワード単位で用いて文書の価値判断をしているので, 意味内容が同じでも, 文書表現によって大いに価値を左右される.

- スコア以外には検索されたテキストに順位をつけることができない

  検索結果を見ると, 現状ではキーワードのマッチ数でしか他の文書と文書の価値に差異をつけることができない. これではたとえユーザの入力したキーワードが文書にあまり含まれていない場合, ユーザの欲しい情報の質, 価値が高いものであっても優先順位は低くなってしまう恐れが可能性がある.

- 検索式やテキスト中の概念間で重要さに差異をつけることができない

  「その文書が何を意味するか」に注目し, その文書の意味内容を理解して, 価値の深さを判断し, 他の文書と概念間で重要さの差異をつけることができれば, ユーザの情報を選択する指標として大きな支えとなる. しかし, 現状の全文検索システムはその文書に含まれているキーワードによる価値判断の違いでしか定義されていず, その文書の意味内容で差異をつけることは不可能である.

- 複雑な論理を使った検索式の作成が求めれる

  本来, 情報を求めるユーザは欲しい情報をうまく言語化できないため, 全文検索システムを使うのだが, 膨大な WWW 上から欲しい情報を獲得するためには, ユーザはあらゆるキーワード, 検索式を定義し, 入力しなければならないという矛盾した問題点が見られる.

その他, 否定演算子 (NOT 検索) で該当するキーワードを含む文書は切り捨てられてしまうという問題点もある.

つまり検索式に必要とされるのは, 文書の表現に使用されているものと同一したキーワードでなければいけない. 例えば, ブール演算子によって適切に構造化された検索式や, 語間になんら関連がないかあるいは漠然とした関連しかない相互に独立なキーワードから構成される検索式である. このような検索式では統語的な統制は行なわれているが, その意味の統制は行なわれていない. すなわち検索式に表されているユーザの問題空間からの潜在的な情報は扱われていない.

第 3 章

全文検索システム「Community Knowledge」] の設計

3.1 本章の概要

本章では第 2 章の全文検索システムの問題点を踏まえ, その問題点を改善し, 本研究における全文検索システム「Community Knowledge」への設計を述べる.

3.2 問題点からの検討

全文検索システムの問題点は大きく分けると 2 つに分けられる.

- ノイズの多い検索結果
- 完全照合手法による問題

3.3 ノイズの多い検索結果への解決

膨大な WWW 上の文書に対して, ユーザがキーワードマッチによる検索をかけると, 検索結果の中に, ユーザの意図とは異なる全く別の内容の文書が混ざる可能性が大きい.

ロボットが WWW 上の文書を回収した後に目的, 趣向別に文書を分類すれば, もしくはある目的, 趣向だけの内容の文書のみ回収すれば, 全く目的, 趣向の異なる文章が混ざるという問題を回避できる.

3.3.1 概念の形成

ユーザは全文検索システムを利用する場合, ある程度の概念が浮かび上がり, はじめて検索行為を行なう. 例えば, 「カズ」について知りたい場合, 「サッカー」という概念から, その中に検索対象を求める.

これはディレクトリ型検索システムを利用して目的の文書を探す要領に似ている.Yahoo![8]や MSN[9] のようなディレクトリ型検索システムは, あらかじめ文書 (WWW ページ) の索引となるキーワード群の構造体を持っている. このキーワード群は一般概念に基づいてツリー構造になっており, 分類された各ディレクトリの内容に応じて WWW ページが登録されていく. 登録の形式については運営主体によって異なるが, 基本的な形式は, WWW ページの提供者が WWW ページの内容に合致すると思われる分類に当てはめていくという形式である.

このような形式を利用して, 膨大な WWW 上の文書からユーザの浮かび上がる概念の意味内容に検索対象を狭めれば, ユーザが意図した文書とは関係ない文書を除外できる. そのためには WWW 上の文書からユーザの浮かび上がる概念と意味内容が一致したページだけを抽出する必要がある.

3.4 完全照合手法への解決

完全照合手法には以下の問題点が挙げられた.

- 適合する文書であっても, その表現が検索式と部分的にしか一致しない場合は, 検索されない
- スコア以外の指標では, 検索されたテキスト間の順位をつけることができない
- 検索式やテキスト中の概念間で, 重要さを基にした差異をつけることができない
- 複雑な論理を使った検索式の作成が求めれる

3.4.1 文書の構造に着目した検索手法

既存の WWW 上の全文検索システムは通常, 文書に使われている自然言語表現に対して, 単語単位で着目している. しかし既存の WWW 上の全文検索システムの索引手法では, 文書中の自然言語表現はすべての単語の頻度を同等に扱っている, という問題点が解決されていない. つまり検索結果としてユーザに返される文書を, 適合性の度合いに関して統制することができない. 高い適合性を維持する方法として, 文書の構造に着目する. 文書の構造を抽出するには, 具体的に単語と単語の類似度を測り, 単語がどのような場面で意味をもって使用されているかを調査する. 単語の共出現を考えることによって, 単語間の意味的なつながりすなわち文書の構造を見い出せる.

3.4.2 検索結果表示の改善

WWW 上の全文検索システムの検索結果の表示を改善し, 索引付けの際に連想辞書やシソーラス展開を行なって, ユーザにその文書に含まれる関連語を提示すれば, 情報を求めるユーザはどの文書が自分の欲しい情報かが把握でき, 文書の選択の指標となる. このような

仲介機能を備えることによってユーザは複雑な検索式を組ませなくても適当に思いついたキーワードを入力するだけでよい.

シソーラスとはあるキーワードについて関連する語を記述したものである.

3.5　全文検索システム「Community Knowledge」の設計

本研究では, 一本化されたユーザの目的, 趣向の意味内容の文書のみを回収し, 次に全文検索システムの仲介機能を改善し, 提供する文書とユーザの要求とが適合するために以下の設計を行なった.

- 文書の概念形成
- 複雑な検索式を必要としない単一キーワード入力のみ
- 従来のスコア以外にも文書の意味内容 (構造) から順位をつける
- 検索式やテキスト中の概念間で重要さに差異をつける

3.6　全文検索システム「Community Knowledge」の処理の流れ

本研究における全文検索システムの処理は, 次のようなルーチンとなる.

1. 概念の形成
2. シソーラスの作成
3. 単語間の関連性の抽出
4. 回収した文書のフィルタリング
5. シソーラス展開による文書のグループ分け
6. 検索結果におけるシソーラスの提示と意味内容による差異づけ

3.6.1　概念の形成

概念を形成するためには, 全文検索システムが「何を目的にするか, 何を対象とするか」を明確化する必要性がある. そうすることによって, 広い検索対象から特定の内容を目的とした文書の抽出に踏み出すことができる.

3.6.2 シソーラスの作成

シソーラスを作成するにはまず概念に沿った単語群つまり単語辞書を作成しなければならない.

単語辞書には,

- 日本語単語辞書
- 英語単語辞書
- 概念辞書
- 概念辞書
- 専門用語辞書

などが存在する.

しかし動的に発信されるWWW上の文書に使われる単語は新しい時代のニーズによって生成された単語である. 例えば,「サッカー」なら「Jリーグ」,「Vゴール」などが挙げられる. そういった動的に生成された単語郡を辞書から抽出するのは不可能であり, 独自に開発することが必須となる.

単語辞書を作成するリーソスすなわちその単語群が掲載されている文書から, 単語を抽出すれば, 単語辞書の作成することは可能である.

ここでは単語辞書を形成する源として, ネットワークコミュニティで流通しているWWWページ, メーリングリスト, ニュースグループなどを利用する.

ネットワークコミュニティの特徴として, 静的・定型的な情報に加え, メンバーの情報活動（調べる, 質問する, 人を知る, 議論する）によって動的な情報が生成されることが挙げられる. この動的な情報は専門性が高く, 有益であり, 情報を発見・理解・整理するための知識の源となる. こういった特徴を生かし, 時代と合わさった単語辞書を形成することが可能となる.

WWWページ, メーリングリスト, ニュースグループなどのデータをネットワークコミュニティ全体における情報活動として記録する. 静的な情報活動記録としてはWWWページに注目し, 動的な情報活動記録としてはメーリングリストとニュースグループのアーカイブに注目する. 情報活動の記録から形態素解析システムを用いて, キーワードの頻度を調査する. 出現頻度の高いキーワードはコミュニティの活発に行われている情報活動=知識のテーマであると考えられる. 抽出した単語から辞書を形成する. ここではこの単語辞書を「コミュニティ辞書」と名付ける.

3.6.3 単語間の類似性を抽出

コミュニティ辞書作成時に単語を抽出すると共に, ネットワークコミュニティによる情報活動の記録から単語の類似性を抽出する. ここではそのコミュニティで使用されている重要

なキーワードのみに注目する. 文書中に使われている単語の頻度には注目せず, 関連性のみに焦点をあてる. 単語と単語の共存度が各文書に抽出できればできるほど, その単語と単語の類似性は高いと断定する. 図 3.1のような類似性を調査した後, それをもとに辞書を作成する. ここではこの単語の類似性を抽出したものを「単語関連辞書」と名付ける.

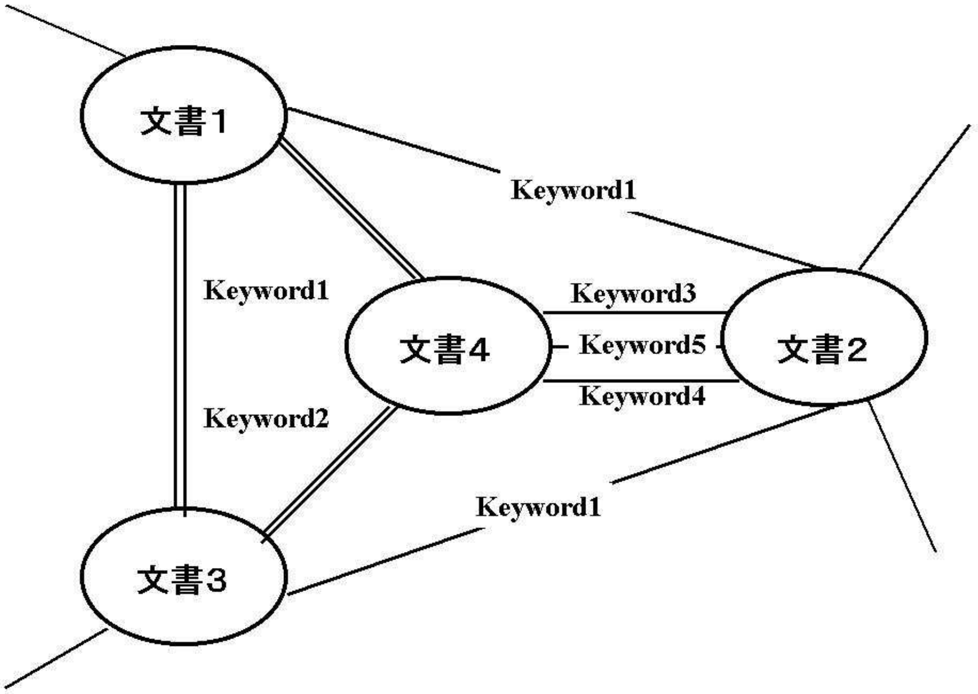

図 3.1: 単語の類似性

3.6.4 回収した文書の概念によるフィルタリング

無作為に回収した膨大な WWW 上の文書から目的化された概念に沿った内容の文書を抽出するためには, その文書が「何を意味するか」を把握し, 分析を行なわなければならない. そうするには, 文書で使用されている全単語群を調査し, 例えば「サッカー」に関する内容を抽出したいのであれば,「サッカー」に関連するコミュニティ辞書と照らし合わせ, その結果これを含むのなら, その文書の内容が「サッカー」を意味することだとわかる. 逆にコミュニティ辞書に掲載された単語が抽出されなければ, その文書はその概念とは関係ない文書とされ, 切り捨てられる.

3.6.5 シソーラス展開による文書のグループ分け

フィルタリングされた文書を単語単位でインデックスすると同時に, コミュニティ辞書を用いてグループ分けを行なう. 各グループの単語が含まれる文書は, 図 3.2のようにそのグループに所属することになる. 各文書は複数のグループに所属し得る. 例えば,「J リーグ」という単語が含まれれば, この文書を「J リーグ」のグループに含める.

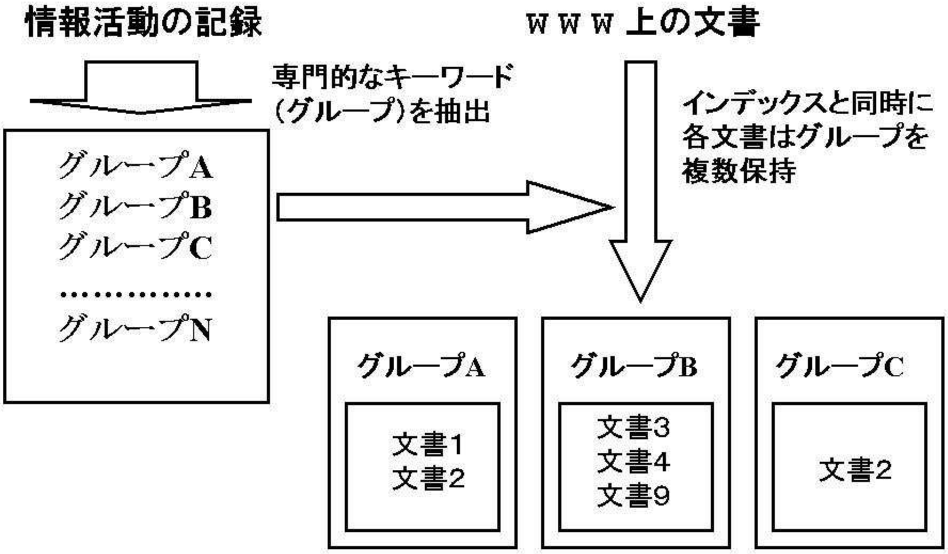

図 3.2: 文書群のグループ分け

3.6.6 検索結果におけるシソーラスの提示と意味内容による差異づけ

ユーザが検索時に自由に単一のキーワードを入力すると, 全文検索システムはそのキーワードが含まれる文書のタイトル,URL などを表示すると同時に, 図 3.3のように選択された文書が所有しうるグループを表示する. その文書が複数のグループに所属していれば, グループ表示も複数とする. また検索結果として返る各文書の順位づけのためにユーザが入力したキーワードから単語関連辞書をみて, 関連するキーワードを求める. 求められた総キーワードが多く含まれている文書を高いスコアとして, 他の文書と差異をつける. これはユーザの意図したキーワードをどれだけ重点に文書中に含まれているかを比較するためである.

しかし, それでも内容の異なる文書で同じ順位の場合もあり得る. このときユーザは検索結果が返ったときに提示されたグループ (シソーラス) を参考にすることによって, 各文書の選択の指標を得ることができ, その結果, 自分の欲しい情報がどれなのかを把握することが可能となる.

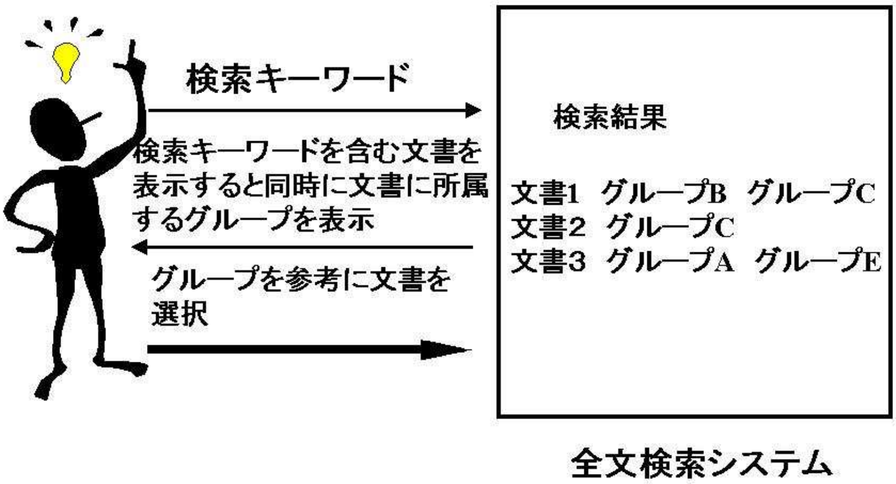
検索キーワード
検索キーワードを含む文書を表示すると同時に文書に所属するグループを表示
グループを参考に文書を選択
検索結果
文書1 グループB グループC
文書2 グループC
文書3 グループA グループE
全文検索システム

図 3.3: グループの提示

第 4 章

システムの実装

4.1 本章の概要

本章では, 前章で設計した全文検索システム 「Community Knowledge」の方針に従い, プロトタイプの実装について述べる.

4.2 環境

今回の実装環境として PC-AT 互換機 (Pentium Pro200,Memory64M) 上で, 以下の OS, ソフトウェアを使用した.

- OS FreeBSD3.0-RELEASE
- WWW Server Apache 1.3.3
- プログラム言語 Perl 5.005_02
- 形態素解析システム すもも

4.2.1 形態素解析システム 「すもも」について

形態素解析システム 「すもも」は標準の辞書として (株) 日本電子化辞書の日本語単語辞書を使用している.

「すもも」の日本語単語辞書は日本語単語辞書レコードを単語見出しの読み順 (五十音順) に並べたものである. 日本語単語辞書レコードは, レコード番号と, 見出し情報, 文法情報, 意味情報, 運用 ·その他情報, および管理情報から構成される. 日本語単語辞書の基本的な役割は, 日本語単語と概念の対応関係を記述し, この対応関係が成り立つときの文法的特性を与えることである。日本語単語辞書レコードは, 日常一般の基本語を対象とする。

4.3 システムの実装

本研究での全文検索システムの実装を以下の順序で説明する.

1. ネットワークコミュニティの情報活動の記録を回収
2. コミュニティ辞書の作成
3. 単語関連辞書の作成
4. WWW 上の文書を回収 (ロボット部)
5. 回収した文書の概念によるフィルタリング
6. 文書のグループ分け
7. 検索処理

4.3.1 ネットワークコミュニティの情報活動の記録を回収

今回対象としたネットワークコミュニティは実在する「障害者関連組織」を対象とする. 用いた文書としては WWW 上の文書 (WWW ページ), メーリングリストおよびニュースグループのアーカイブである. 各記録の詳細は以下の通りである.

- WWW 上の文書　「月刊 全国障害者介護制度情報」http://www.top.or.jp/%7Epp/
- メーリングリスト　「身障者生活情報メーリングリスト (disabled ML)」
- ニュースグループ　「fj.misc.handicap」

今回は WWW 上の文書を 200 個, メーリングリストのアーカイブを 240 通, ニュースグループのアーカイブを 300 通回収した.WWW 上の文書を回収するためのロボットとして GNU Wget[14] を採用した.

4.3.2 コミュニティ辞書の作成

収集した情報活動の記録群 (WWW ページ, メーリングリストのアーカイブ, ニュースグループのアーカイブ) を, 所定のディレクトリに置く. この記録群がコミュニティ辞書を作成するリソースとなる. この文書群からコミュニティ辞書作成プログラムを利用し, ネットワークコミュニティがよく使用している単語とその頻度を抽出し, それをファイル keyword_index に格納した.

より具体的には, 図 4.1のように各文書に対して形態素解析プログラム「すもも」を使用して, 単語分けを行ない, 固有名詞, 普通名詞, サ変名詞, 人名, 未定義のみを抽出した. 使用頻度の高い単語を抽出するため, 抽出する単語の条件としては,2 つ以上の文書で使用され

ている単語のみに注目した. 単語の中には一般的な名詞が含まれているが, 使用頻度が高い単語は該当するネットワークコミュニティではよく使用される重要な単語とみなした.

しかし使用頻度のみに着目すると, コミュニティ単位でなく, 名前や住所などの情報提供者単位で使用頻度の高い単語を抽出する恐れがあるので, 今回は異なる情報提供者の記録で使用頻度が 2 以上の単語を抽出した. その他, 図 4.2のように数字や半角カナ使用の単語は除外した.

```
% perl5 mk_group.pl -d files -f 2 -g

Directory: files

Read file: 19980106-01
Temporary Directory: /tmp/files
Temporary File     : /tmp/files/19980106-01
 cutting header files/19980106-01
 saving /tmp/files/19980106-01
Word:ポスト Attribute :サ変名詞
Word:ＪＲ Attribute :固有名詞
Word:京都 Attribute :固有名詞
Word:駅 Attribute :普通名詞
Word:駅員 Attribute :普通名詞
Word:遭遇 Attribute :サ変名詞
Word:新幹線 Attribute :普通名詞
Word:ＪＲ Attribute :固有名詞
Word:奈良 Attribute :固有名詞
Word:ＪＲ Attribute :固有名詞
Word:奈良 Attribute :固有名詞
Word:車椅子 Attribute :普通名詞
Word:駅名 Attribute :普通名詞
```

図 4.1: コミュニティ辞書の作成

```
  while (<FILE>) {
    chomp;

    /(\S+)\s+(\S+)/;
    $word = $1;
    $attr = $2;

    ## select keyword
    next unless ($attr =~ /固有名詞/ or $attr =~ /普通名詞/ or
                 $attr =~ /サ変名詞/ or $attr =~ /未定義語/);
    next if $word =~ /^\d/;
    next if $word =~ /^_/;
#    next if $word =~ /^\w/;
    unless ($word =~ /^\w/) {
      next unless ($word =~ /^[\241-\376][\241-\376]/ or
                   $word =~ /^[\217][\241-\376][\241-\376]/);
    }
    $word =~ tr/[A-Z]/[a-z]/ if $word =~ /^\w/;

    print "Word:$word Attribute :$attr\n" if $debug;

    $keyword_frequency{$word}++;
    $keyword_tmp{$word}++;

  }
  close FILE;
```

図 4.2: 単語の抽出

抽出された単語	頻度
障害	116
キー	94
情報	83
人	76
会	70
ページ	61
車椅子	61
電話	56
参加	54
ポスト	51
利用	51
ホーム	49
駅	48
片手	46
キーボード	44
投稿	42
日本	40
福祉	36
パソコン	33
ブラインド	33
市	33

表 4.1: コミュニティ辞書

4.3.3 単語関連辞書の作成

ネットワーク上のコミュニティの情報活動の記録群から単語を抽出し, コミュニティ辞書を形成すると同時に, 単語関連辞書の作成を行う. ある文書から抽出した各キーワードはすべて互いのキーワードと関連性が 1 つあると定義する. 今回は各文書間での関連性を求めるので, 同じ文書に 2 以上の関連が抽出されても関連は 1 にする.

品詞別で重要な単語を抽出する方法だと, 「車椅子」などの障害者関連組織にとっては重要な単語は、品詞から見ると, 「普通名詞」になる. 一般的な単語でもコミュニティから見れば, 重要な単語となり得る単語が存在する中で, 品詞による単語の重要性を測ることは不可能である. よって今回は関連度の高さでその単語の重要性を測った. しかしコミュニティ辞書作成と同様に, 名前や所属など情報提供者単位における文章に特化した単語は除外した.

各文書群から関連性を抽出し, 各単語をキーに, その単語と関連のある単語と関連度を求めた上で、表 4.3.3ファイル keyword_relation に格納し, 単語関連辞書を作成した.

単語	関連した単語とその頻度
いす	車,6 駅,4 操作,4 車椅子,4 駅員,3 エスカレータ,3
エスカレータ	宿駅,3 駅,3 中央線,3 設備,3 設置,3 ホーム,3 いす,3
エレベーター	新幹線,3 人,3 駅,3
オリンピック	パラリンピック,3 視聴率,3 放送,3 差別,3
論文	研究,3
サービス	福祉,4 情報,4 会社,4 日本,3 内容,3 予約,3 利用,3
マシン	パソコン,4 配置,3 キーボード,3 キー,3

表 4.2: 単語関連辞書

4.3.4 WWW 上の文書を回収 (ロボット部)

次に, 検索対象となる WWW 上の文書を回収する. 今回は無作為に意味内容の異なる 520 個の文書を回収した.

4.3.5 回収した文書の概念によるフィルタリング

回収した文書から単語を抽出し, コミュニティ辞書に掲載されている単語と比較する. 文書から抽出した単語とコミュニティ辞書に掲載されている単語に一つも該当しない文書は文書集の意味内容と概念 (ここでは障害者関連) には関連性がないとみなし, 検索対象から除外する. これにより, 目的化された概念とは関係のない文書は検索結果としてユーザに返されないようにする.

4.3.6 文書のグループ分け

回収した文書をフィルタリングした後, コミュニティ辞書に掲載されている単語からグループ分けを行なう.

コミュニティ辞書に掲載されている単語をグループとして, 該当する単語が含む文書はそのグループに属することにする. グループ分け作業を行なったあと, ファイル Group_index に格納した.

グループ分けと同時に, 単語単位でのインデックスの作成も行なう. インデックスファイルには, 単語をキーとし, 単語を含んだ文書のポインター情報を格納した Key_index と文書をキーとし、文書に含まれる単語群を格納した File_index を用意した.

4.3.7　検索処理

検索クライアントは今回,Netscape Navigator や Internet Explorer などの WWW ブラウザをそのまま使用する. ユーザは WWW ブラウザを使って, 全文検索システムの入力画面フォームにアクセスする. 検索処理自体は CGI(Common Gateway Interface) によって作動される.

この検索処理は提案に基づき, 次の処理を行なうよう実装した.

1. 検索キーワード入力

   ユーザは, 入力画面から自分の欲しい文書に含まれていると思われるキーワードを入力する.

2. インデックスファイルを参照

   ユーザが入力したキーワードを元に全文検索システムは、インデックスファイル Key_index を参照する.Key_index はキーワードがキーとなっているので、キーワードから対象文書へのポインタを得る.

3. 関連キーワードによる順位づけ

   ユーザの入力されたキーワードに関連するキーワードを, keyword_relation から参照する. 関連するキーワードを抽出した後,File_index と文書の含むキーワードと関連するキーワードを比較し, 各文書が含む関連キーワードの総数を計算する. 各文書の関連キーワードの総数を比較し, 一番総数の多いものから順に順位づけを行なう.

4. 対象文書のグループ情報を参照

   Key_index から得た文書のポインタ情報から Group_index による各文書が所属するグループ情報を取得する.

5. 検索結果として表示

   順位づけされた文書から図 4.3のように順に, タイトル, 総関連キーワード, グループ情報,URL を提示する.

4.3.8　検索結果の例

図 4.3は, 検索キーワード「車椅子」を入力し, 検索結果として返った例である. 上から順にタイトル, 関連キーワード, グループ情報,URL が表示されている.

d.jp/~katsura/prog/search.cgi?search_word=%BC%D6%B0%D8%BB%D2&search_type=and&show_keyword=10

車椅子

検索結果 8 件

スペースに続けてキーワードを追加することにより絞り込検索ができます。

1. 車椅子利用者のための駅情報(メインメニュー)
関連キーワードとスコア: 私 9 駅 8 利用 8 ホーム 6 情報 5 利用者 3
グループ: 駅 情報 利用者 版 私 ページ 利用 データ 近畿 ホーム
http://www02.so-net.ne.jp/%7Ekonin/station/index.html
1840 bytes

2. JR駅の福祉設備
関連キーワードとスコア: 線 5 障害 5 設備 4 新幹線 4 エスカレーター 3
グループ: 線 福祉 設備 新幹線 ページ サービス エスカレーター 東海 障害 県
http://www.portnet.co.jp/fukushi/jr.html
4158 bytes

3. 車椅子で利用できる公衆電話
関連キーワードとスコア: 利用 8 障害 5 施設 5 設備 4 設置 4
グループ: 福祉 ページ 設備 施設 検索 設置 障害 サービス お願い 利用
http://www.portnet.co.jp/fukushi/telbox.html
2306 bytes

図 4.3: 検索結果

第 5 章

評価

5.1　本章の概要

本章では, 前章で実装を行なった全文検索システム「Community Knowledge」の適合率による評価, 順位付けによる評価とグループ提示の評価を行ない, 考察と分析を述べる.

5.2　適合率の評価

5.2.1　適合率と再現率

全文検索システムの評価をするためには, 2 つの重要なパラメータがある. 「再現率」と「適合率」がそのパラメータとなる. 本システムの評価もこれらのパラメータを用いて, 評価を行う.

再現率 (recall) とは、「ユーザの求める文書がどの程度検索されているかを示す割合」であり, 図 5.1の式で定義される.

$$\text{再現率} = \frac{\text{検索された該当文書数}}{\text{全文書中の該当文書数}}$$

図 5.1: 再現率

一方, 適合率 (precision) とは, 検索された文書中にユーザが求める文書がどれくらいあるかを示す割合であり, 図 5.2の式で定義される.

$$\text{適合率} = \frac{\text{検索された該当文書数}}{\text{検索された文書数}}$$

図 5.2: 適合率

これら 2 つのパラメータの関係を摸式的に表すと, 図 5.3の様になる.

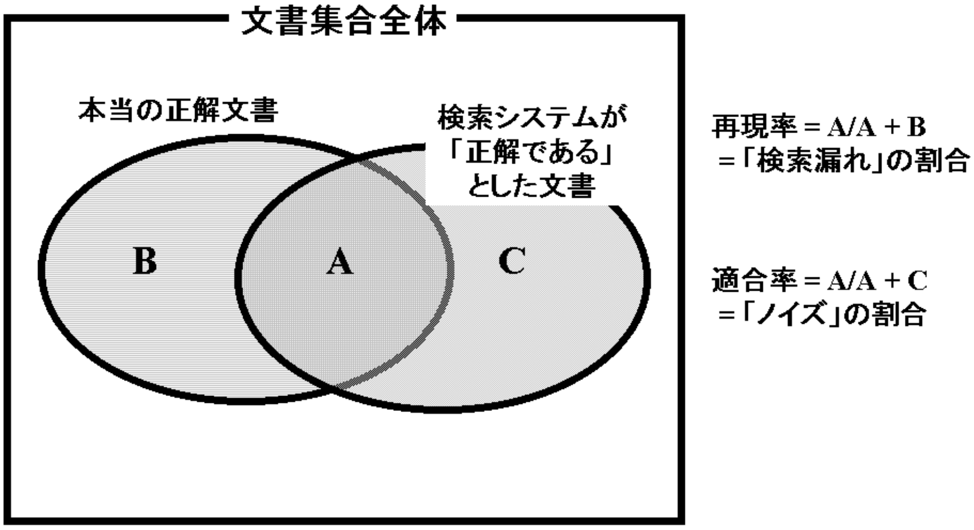

図 5.3: 再現率と適合率の関係

理想的には, これらの 2 つのパラメータの両方が,1 に近付くことが望ましい. つまり,

- 検索洩れがなく (再現率が 1)
- かつ検索された文書には一切ノイズが含まれず
- すべて該当文書である (適合率が 1)

という状態である. しかし実際には, これら 2 つのパラメータの間には, 再現率が上がると, 適合率が下がり, 再現率が下がると適合率が上がるというトレードオフの関係が存在する. いずれのパラメータを重視するかで全文検索システムの特徴が決まるのだが、今回は適合率重視での評価を行い, 「ノイズの多い検索結果」の問題について解決されたかどうかを確証してみた. また適合率による評価を行なう際には, 予め 「正解」となるべき文書を定義しておく必要がある. 文書の意味内容を考慮した全文検索システムでは, このような正解を定義することは容易ではないが, 今回は実際の文書を閲覧して, 著者の主観によって 「正解」かどうかを判断した.

5.2.2 評価

今回は無作為に抽出した WWW 上の文書 500 個と障害者関連情報の文書を 20 個に対して, 「コミュニティ辞書」によるフィルタリングを行い, 適合率によるフィルタリングの効果を測定した. なお適合性の基準は, 著者の主観によりその判断を行っている.

次に障害者生活情報に関連するキーワードを用意し, 検索キーワードの候補として挙げた. 以下が該当する検索キーワードである.

- 車椅子
- 手話

これらのキーワードを基に適合率を測った.

検索キーワード	検索された該当文書数	検索された文書数	適合率
車椅子	7	8	0.875
手話	9	10	0.9

表 5.1: 適合率の評価

表 5.1を見ると, 適合性が高い検索結果が返り, ノイズの多い検索結果から解放された様に見えるが,520 個の文書に対してフィルタリングを行なった際, 除外された文書は 0 個であった. すなわち今回の検索対象はフィルタリングされていない状態と全く変わらなかった. 今回の適合性におけるフィルタリングの評価は期待できないことがわかった. これは 4.3.2 で述べたコミュニティ辞書の作成方法に欠陥があると思われる.

5.3 順位付けの評価

次に順位付けの評価を行なう. 検索の順位づけは, 検索結果の確かさを表すために用いられる. 大量の検索結果を得ることが現在の全文検索システムでは, 順位付けを行って優先順位を並べ替える.

全文検索システムのスコアの基本となるのが, 検索キーワードの出現頻度である. 今回の評価としては出現頻度のみのスコアリングと関連キーワードによるスコアリングの評価を

試みる. 全体の文書集合は「CHIME's SQUARE 障害者情報インデックス」にリンクされいている WWW 上の文書 493 個とし, 検索キーワードとしては「手話」を候補とした. 全節でフィルタリング（検索対象の絞り込み）の評価は期待できないものになったので, あらかじめ障害者に関連する文書集合を対象とした. これによって, 意味内容による順位付け, 文書間による概念間の重要さの差異の度合いを評価する. 文書の意味内容の適合性は適合率の評価同様, 著者の主観で判断を行なった.

順位	文書のタイトル	出現頻度によるスコア
1	WWW server list	113
2	ヒューマンインターフェース	14
3	information Center about DEAF	12
4	茨城県手話サ連	11
5	手話通訳（者）関連資料などのページ	6
6	japanese resources	6
7	prefectural resources	6
8	手話ワンポイントレッスンコーナー	6
9	Japanese Association of Sign Liguisitics	6
10	ユニバーサル端末研究室ホームページ	3

表 5.2: 出現頻度のみのスコアリング

順位	文書のタイトル	関連キーワードによるスコア
1	japanese resouceces	7
2	WWW server list	6
2	ヒューマンインターフェース	6
2	prefectural resoucrces	6
5	information Center about DEAF	4
5	手話通訳（者）関連資料などのページ	4
7	ユニバーサル端末研究室ホームページ	2
8	茨城県手話サ連	1
8	Japanese Association of Sign Liguisitics	1
8	手話ワンポイントレッスンコーナー	1

表 5.3: 関連キーワードによるスコアリング

表 5.2, 表 5.3のスコアリングを基に文書の内容を確認したところ, 関連キーワードによる同じスコアの文書の内容はさほど変わらなかった. 順位で言えば, 図 5.3の 1,2 位の文書の意味内容は障害者関連のリンク集が掲載されている文書であり, 残りの文書の内容は手話に関した文書の索引ページ (index ページ) であった. つまり出現頻度によるスコアリングは, 適合性の度合いについてなんら統制ができてないことがわかった. それに比べ, 関連キーワー

ドによるスコアリングは意味内容による順位付け, 文書構造による重要さの差異をつけられることにより, 適合性の度合いについて統制が可能となった.

5.4 グループ提示の評価

最後にグループ提示の評価を行なう. 全文検索システム「Community Knowledge」によるグループの提示は検索結果が返った時のユーザの各文書の選択の指標となる. 今回は評価環境として, 文書集合は「CHIME's SQUARE 障害者情報インデックス」にリンクされいているWWW上の文書493個とし, 検索キーワード「バリアフリー」を候補とし, 検索処理を行なう. 検索結果に提示されるグループの提示の正確さを確かめるために著者の主観で判断し, 評価を行なった.

文書のタイトル	グループ
IWE'96 高知県/バリアフリーについて	文字, テキスト, 人, 障害, 点字, 県
バリアフリー研究会	研究会, ページ, 支援, 障害, 環境, 社会
japanese resources	障害, ページ, 福祉, 会, ホーム, 情報, センター
Vision Impairments	障害, 視覚, 点字, 日本, 情報, 協会, 盲学校
prefectural resources	障害, ページ, 会, 福祉, ホーム, 情報, センター
「障害者支援情報機器システムに関する調査」結果	調査, 機器, 福祉, 支援, システム, 開発, 障害
IH 調理器_ICBF1_ホームページ	テキスト, 社会, 紹介, 使用, 発売
CHARIER	番号, ビル, 株式会社, 段差, 巻, 段差, 設計,
WWW server list	手話, 障害, 聴覚, 大学, 会, 情報, 協会, 学校

表5.4: グループの提示

表5.4が検索結果として返った各文書のタイトルとグループである. 表5.4にあるグループを参考に文書の内容を確認したところ, 提示されたグループが文書の意味内容に深く関係していることが確認できた. 例えば, 表5.4の上の「IWE'96 高知県/バリアフリーについて」の文書の内容は, 高知県による視覚障害者のためのWWWページのテキストについてであるが, グループの提示である「文字, テキスト, 人, 障害, 点字, 県」と文書の意味内容が合致していることがわかった.

5.5 考察

今回はユーザは単一キーワードによる検索で, 全文検索システムによる関連キーワードによる順位付け, 文書構造(グループ)の提示により, 全文検索システム「Community Knowledge」の優位性を証明した. しかし検索対象のフィルタリングは評価の通り, 失敗に終ってしまった。この原因となるコミュニティ辞書の欠陥を次節の分析で示したいと思う.

5.6 分析

コミュニティ辞書の欠陥を証明するために, 著者の主観により, 障害者関連の専門性の高いと思われる単語を抽出し, 表5.5の疑似コミュニティ辞書を作成した.

抽出された単語	頻度
障害	116
情報	83
車椅子	61
電話	56
投稿	42
福祉	36
インターネット	21
コンピューター	18
ボランティア	16
手話	15
難聴	13
団体	13
身体	13
パラリンピック	12
点字	10
障害児	9
心臓	9
リハビリテーション	8
シンポジウム	7
半身不随	7
聴覚	7
バリアフリー	7
身障者	7
電磁波	7
視覚	6

表5.5: 分析用のコミュニティ辞書

分析用のコミュニティ辞書5.5を用いて適合率の評価同様, 無作為に抽出した500個のWWW上の文書と障害者関連情報の文書の20個に対してフィルタリングを行なった.

その結果, 対象とした520の文書のうち,476個の文書が振り落され,44個の文書が絞り込まれた. サンプルとして混入させた20個の障害者関連情報の文書は全てこの44個の中に含まれていた.

コミュニティ辞書がコミュニティの知識を適切に反映している場合には, このフィルタリングは有効であることが示された. しかし実際のフィルタリングのパフォーマンスはコミュニティ辞書の内容に依存するため, コミュニティの知識を適切に抽出する手法が今後の課題となる.

回収した文書の数	フィルタリングされた文書数
520	44

表 5.6: 分析結果

表 5.6のフィルタリングされた文書数において, 適合率の評価を確かめると, 表 5.1の適合率が得られた. よってコミュニティ辞書作成時における単語の抽出方法に欠陥があることがわかった. コミュニティ辞書作成方法の改善を今後の課題にしたいと思う.

第 6 章

結論

6.1　まとめ

本研究では全文検索システム『Community Knowledge』の設計・実装・評価を行なった. 既存の WWW 上の全文検索システムにおけるノイズの多い検索結果, 完全照合手法による問題に取り組んだ. ノイズの多い検索結果の解決方法として, コミュニティの情報活動の記録を利用して, 検索対象の文書集合を絞りこんだ. また完全照合手法の解決方法としては, 文書の自然言語に対して, 単語単位でなく, 文書構造に着目し, 文書の意味内容に差異をつけ, 検索結果にシソーラスを提示した.

評価として設計・実装した全文検索システム『Community Knowledge』の検索結果を閲覧し, 検索結果として返される文書の意味内容を確認し, 文書の順位づけとグループ情報の提示において, 優位性を持つことを示した.

コミュニティの知識を反映させることでユーザが自分の欲しい情報を手にすることができる全文検索システムを設計・実装できた.

6.2　今後の課題

ネットワークコミュニティの知識を反映させた全文検索システム『Community Knowledge』を更に改善するためには, 以下の課題が挙げられる.

- コミュニティ辞書の作成手法
- 様々なコミュニティにおける試行評価 (汎用性)
- 検索洩れへの対応
- 客観的な評価

今回, コミュニティ辞書による検索対象のフィルタリングは良い評価が得られなかった. これは第 5 章の評価と分析にも述べたように, その作成方法に欠陥があった. この欠陥を修正し, 検索対象をうまく絞り込めるコミュニティ辞書の作成を課題としていきたい.

また今回は「障害者関連組織」に関する概念に限定したため, 他の概念による試行評価を行なっていず, 汎用性を確かめていない. 今後は様々なコミュニティの知識を反映させた全文検索システムを設置したいと思う.

また, 再現率を考慮せず, 検索洩れへの対応も未解決のままである. 検索キーワードを含まずに検索キーワードと意味内容が適合する文書が検索結果として返るような設計を目指さなければならない.

今回の評価は著者の主観による評価のみであったので, 今後は客観的な評価基準から, コミュニティ知識を反映させた全文検索システムを評価し, その優位性を証明していきたい.

このような課題を取り組み, ユーザにとってより使いやすい全文検索システムの設計と実装を進めていきたいと思う.

6.3 おわりに

全文検索技術の研究は様々なところで行なわれている. しかしこれは, 基礎的研究が終った成熟な分野であることを意味するものでは決してない. 多くのユーザ層からの要求に応え得るだけの必要充分な機能を提供しているわけではない. 幅広い要求に応えるためにはさらなる技術革新が常に求められている.

しかし本研究での全文検索システムの手法はユーザの要求の一部しか満たしておらず, 従って様々な問題点が残されている. そのためには, 従来の自然言語学および情報処理学の分野のみならず, 認知科学, 人工知能, 社会学, 心理学などさまざまな分野にわたる分野横断的な研究を行なう必要がある.

こういった研究成果がフィードバックされていくことによって, 今後より高度な方法をより簡便な方法で行なうことが理想の全文検索システムのスタイルになっていくと予想される.

謝辞

本研究を進めるにあたり, 多大なる御助言と御指導を頂きました, 慶應義塾大学環境情報学部教授の村井純博士, 徳田英幸博士に感謝します. また, 絶えず御助言と御指導頂きました, 同学部の中村修博士, 楠本博之博士に深い感謝の意を表します.

また本論文の執筆に際し, 研究に関する情報, 論文の書き方, 発表の手順など多くの助言を与えて下さった慶應義塾大学大学院政策・メディア研究科博士課程の宮川祥子氏, 論文提出の前日に徹夜で論文の修正につき合って頂いた同大学大学院政策・メディア科修士課程の大西孝義氏, 余語徹朗氏, いつも精神的に私をサポートして頂いた石田剛明氏にも感謝します.

そして同大学環境情報学部伊藤俊介氏には試験中で忙しい中, 英語が苦手な私に惜しみない協力を頂き, 感謝します.

その他, 共に研究活動を行ない, 様々な助言を与えて下さった慶應義塾大学徳田・村井研究室の諸氏に感謝します.

以上をもって謝辞と致します.

参考文献

[1] 馬場 肇: 「日本語全文検索システムの構築と活用」,SOFTBANK,1998

[2] Peter Ingwersen: 「情報検索研究 -認知的アプローチ-」, トッパン,1995

[3] Joseph P.Bigus 「ニューラル・ネットワークによるデータマイニング」, 日経BP社,1997

[4] Larry Wall: Programing Perl 2nd Edition,O'Reilly & Associates,Inc.,1996

[5] Ingrid: http://www.ingrid.org/papers/RDpaper.html

[6] RCAAU(MONDOU): http://www.kuamp.kyoot-u.ac.jp/labs/infocom/mondaou/

[7] Altavista: http://www.altavista.com/

[8] Yahoo!: http://www.yahoo.com/

[9] MSN: http://www.msn.com/

[10] Lycos!: http://www.lycos.co.jp/

[11] Infoseek: http://www.infoseek.com/

[12] ODIN: http://odin.ingrid.org/

[13] Excite: http://www.excite.com/

[14] Wget: http://www.gnu.org/software/wget/wget.html

[15] JUMAN: http://casper.ish.ic.kanagawa-it.ac.jp/ okamoto/

[16] JENOMI: 図書館大学

[17] BreakFast: http://www.fujitsu.co.jp/hypertext/free/breakfast/download.html

[18] SuperMorpho-J: http://www.omron.co.jp/supermorpho-j.html

[19] 茶栓: http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[20] すもも: http://www.brl.ntt.co.jp/sumomo