平成 12 年度

学士学位論文

類似事例データを活用した
ログ解析システムの構築に関する研究

A Study on Construction of a Log Analysis System Utilizing Similar Case Data

1010418　　友池 貴之

指導教員　　Ruck Thawonmas 助教授

2000 年 2 月 5 日

高知工科大学 情報システム工学科

要 旨

類似事例データを活用した
ログ解析システムの構築に関する研究

友池 貴之

本論文では, ラック研究室で開発された高知工科大学ヘルプシステムにおいて, ログを自動的に解析するログ解析システムの構築と, その解析の結果について述べる. 本システムは, 保守者による目視の解析結果を事例データとしてデータベース化し, その後のログを自動的に処理することを基本とする. 本システムは, 事例がまったく無い状態から解析をはじめ, 類似の条件を厳しくした場合でも約 13.6%のログに対して自動的に解析を行うことができている.

キーワード　　自然言語, 情報検索, ベクトル空間法, 形態素解析

Abstract

A Study on Construction of a Log Analysis System Utilizing Similar Case Data

Takayuki TOMOIKE

This thesis describes construction of a Log Analysis System that automatically analyzes logs in the Kochi University of Technology Help System developed at the Ruck Laboratory. The system is based on analysis results viewed by the system operator. These results are considered case data and then stored in database for automatically processing subsequent logs. Even if the system begins analysis from the state without any case data, and even when the similarity testing condition is made severe, the system can automatically analyze logs of up to about 13.6%.

key words Natural Language, Information Retrieval, Vector-Space Model, Morphological Analysis

目次

図目次

表目次

第 1 章

はじめに

WWW の普及にともないネット上で発信される情報が急増している [1]. 今や情報を検索・閲覧する手段として WWW は生活に欠かせないものとなりつつある. しかし, 発信されているそれらの情報の質・量に比べて, 知りたい情報を素早く的確に検索・閲覧する手段は, まだ十分とはいえない状況である.

現在一般的に使用されている情報検索は, 知りたい情報のキーワードを利用者が指定する方式である. 検索サービスの構築が容易, ソフトウェアの原理が簡単などの理由で広く行われているが, 利用者が自在に利用できるものにはなっていない.

本論文では, まず, この問題を克服し試験運用を開始した自動応答システムについて述べる.

本論文で述べる自動応答システムは, 株式会社 エス・エス・アール[*1], 高知工科大学 情報システム工学科 坂本研究室, 同ラック研究室の共同研究の成果によるものである. 同システムは, 自然言語で書かれた質問に対して適切な答えを返すことを特徴としている.

次に, 本論文で提案する自動応答システムのログを解析するシステムについて述べる.

自動応答システムをはじめとする情報検索システムにおいて, 検索結果のログについて, 妥当性, 有効性などを解析することは, 情報検索システムの処理性能や検索に関する質の向上にとって極めて重要である [2].

従来は検索結果のログを保守者の目視により調査・評価する方法が採られていた. しかし, この方法では, 保守者への負担が大きい, 人為的なミスを排除できない, 保守者の判断が曖昧

[*1] http://www.e-ssr.co.jp/

であり評価に一貫性が無いなどの弱点があった.

本論文で提案するログ解析システムは, 担当者による目視の評価を事例データとしてデータベース化し, その後に発生した検索結果のログを自動的に処理するインテリジェントなシステムである.

第 2 章

自動応答システムの構築

自動応答システムとは, よく聞かれる質問とその回答をデータベース化し知識ベースとして蓄え, コンピュータが自動的に質問を受け付け, 回答する情報検索システム [3] の一種である.

現在, インターネット上で一般的に使用されている情報検索システムは, キーワードを利用者が指定する方式である. この方式は, 検索サービスの構築が容易, ソフトウェアの原理が簡単などの理由で広く行われている. しかし, このキーワード型の情報検索方式は, 利用者の使い勝手を考慮しているとは言い難い.

たとえば, ひとつだけのキーワードでは検索がうまく行えない場合, キーワードを複数指定して検索することになる. このとき, 利用者には AND 条件／ OR 条件などといったソフトウェア的な知識が必要となる. この方式では, 素早く・的確に知りたい情報を取り出すという利用者の要求を十分に満たしているとは言い難い.

本論文で述べる自動応答システムは, 自然言語で書かれた質問に対して回答できることを大きな特徴としている. これにより, 利用者はまるで人間に対して質問をする感覚で使用することができ, 知りたい情報を正確に特定できる. また, 言い回しの違う質問や類似語による質問に対しても同様に回答できる.

さらに, 知識ベースを自然言語で記述できることにも特徴を有する. これは, システムの保守・拡張を容易にするものである.

自動応答システムの応用として, 図 2.1 に示す高知工科大学ヘルプシステムが構築された. 同ヘルプシステムは, 高知工科大学事務局の協力を得て入学を目指す受験生を対象とし, おもに受験に対する疑問・質問に答えるものである. 同ヘルプシステムは, 2000 年 7 月の学内

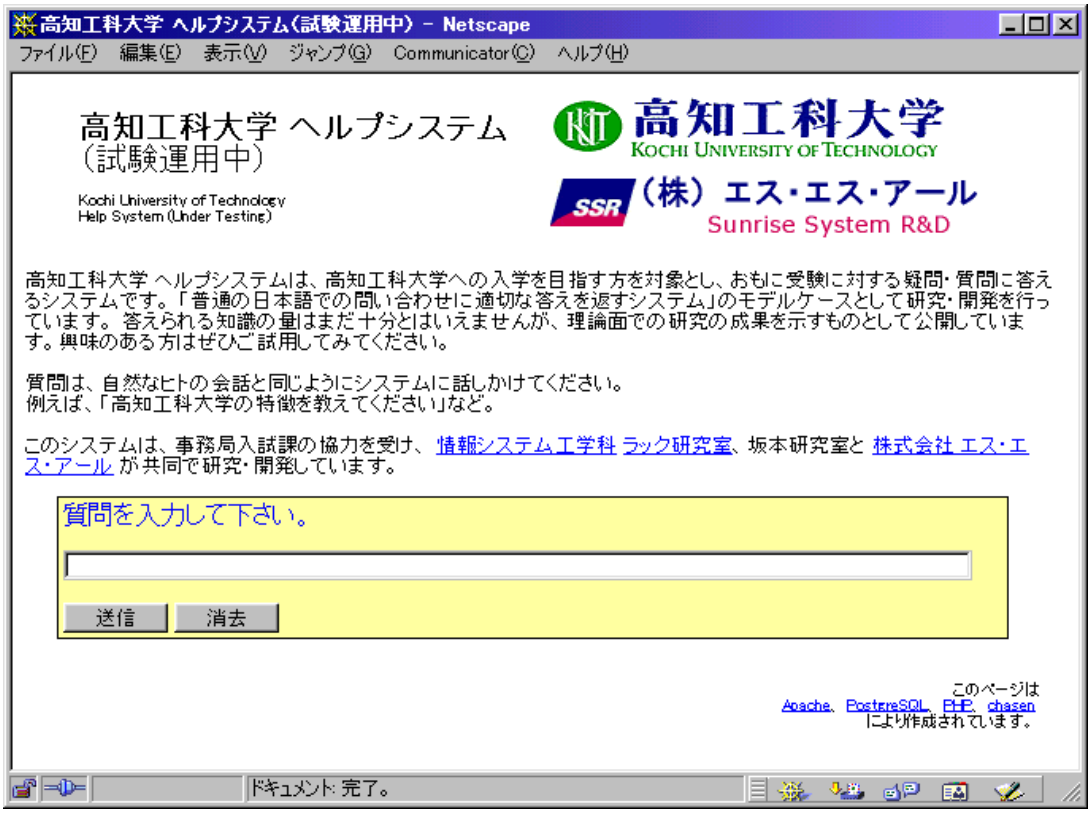


図 2.1　高知工科大学ヘルプシステム

公開の後, 同年 9 月から本学のホームページ[*1]にて試験公開中である.

2.1　質問パターンの分類

従来のキーワード型の検索システムでは, 用意している情報の中から利用者が指定したキーワードを含んでいる情報を探して回答する. これに対して本自動応答システムでは, 利用者に自然言語で書かれたテキストで質問してもらい, 利用者が何を知りたがっているかを判断する.

これを実現するために, 本自動応答システムでは, 想定される質問の言い回しパターンが用意されている. 質問パターンを利用することで, 同自動応答システムは, 言い回しによる質問の違いを理解している.

[*1] http://www.kochi-tech.ac.jp/

2.1 質問パターンの分類

高知工科大学ヘルプシステムを構築するにあたって, まず, 学生が大学に関してどのような質問を持っているのかを調査した. この調査は, 1999 年度セミナー 3 及び 2000 年度セミナー 1 にて情報システム工学科の 3 回生, 1 回生の合わせて約 160 名の学生に質問を書いてもらい集計したものである. その結果, 以下のように質問パターンを分類できることががわかった.

What is 型：内容を問うもの.

高知工科大学の誇れる所を教えてください.

試験科目は何ですか.

Can I 型：可能・不可能を問うもの.

ドミトリーや大学構内は携帯電話の電波は入りますか.

留学はできますか.

Is there 型：存在を問うもの.

近くにアルバイトをする場所はありますか.

奨学金の制度はあるのか.

When 型：時を問うもの.

入試はいつですか.

夏休みはいつからですか.

Where 型：場所を問うもの

インターンシップの受け入れ先はどこですか.

授業で使う教科書はどこで買うのですか.

How much 型：お金の価値を問うもの

授業料はいくらですか.

入学時に費用はいくら必要ですか.

How many 型：数量を問うもの

就職率はどれくらいですか.

学食の座席はいくつありますか.

高知工科大学ヘルプシステムでは, これらの What is 型, Can I 型, Is there 型, When 型, Where 型, How much 型, How many 型を質問パターンとしている. この質問パターンを利用することで, 質問の言い回しによる違いの解消している.

2.2　知識ベース

多量のデータを整理整頓し管理することは, 検索効率を上げるためばかりでなく, システムの保守・拡張を容易にする点で重要な課題である. また, システムの保守・拡張を容易にするという点においては, データは人間が見てわかる形であるのが望ましい.

そこで, 本自動応答システムでは図 2.2 のような自然言語で記述されたデータをフレーム型で管理する方式が採用された.

知識ベースは, 図 2.3 のようにフレームが親子の関係を持ったフレームで構成されている. 各フレームは, それぞれ一つの知識ドメインで構成されている.

フレームは, タイトル, 親フレーム, 基本情報, 質問回答ペアの集まりで構成されている.

タイトルは, 各フレームの名称であり知識ドメインを代表する言葉が当てはまる.

親フレームは, フレームの親子関係を示す情報である.

基本情報は, タイトルの説明であり自動応答機能の性能向上のために使用される.

質問回答ペアは, よく聞かれる, または, 聞かれるであろう質問とその回答のペアである.

このように, 自然言語で書かれたデータをフレームという概念で構造化することにより, コンピュータと人間の相互で利用可能な柔軟な知識ベースを構築することができる.

2.3　マッチング

マッチングを中心の機能としたときの自動応答システムの構成図を図 2.4 に示す.

まず, 本自動応答システムが受け付けた検索要求の質問は, 形態素解析されて形態素に分解される. 形態素解析のツールとして, 計算機による日本語の解析の研究を目指す多くの研究者に共通に使える形態素解析ツールを提供するために開発された 茶筌 [4] を使用して

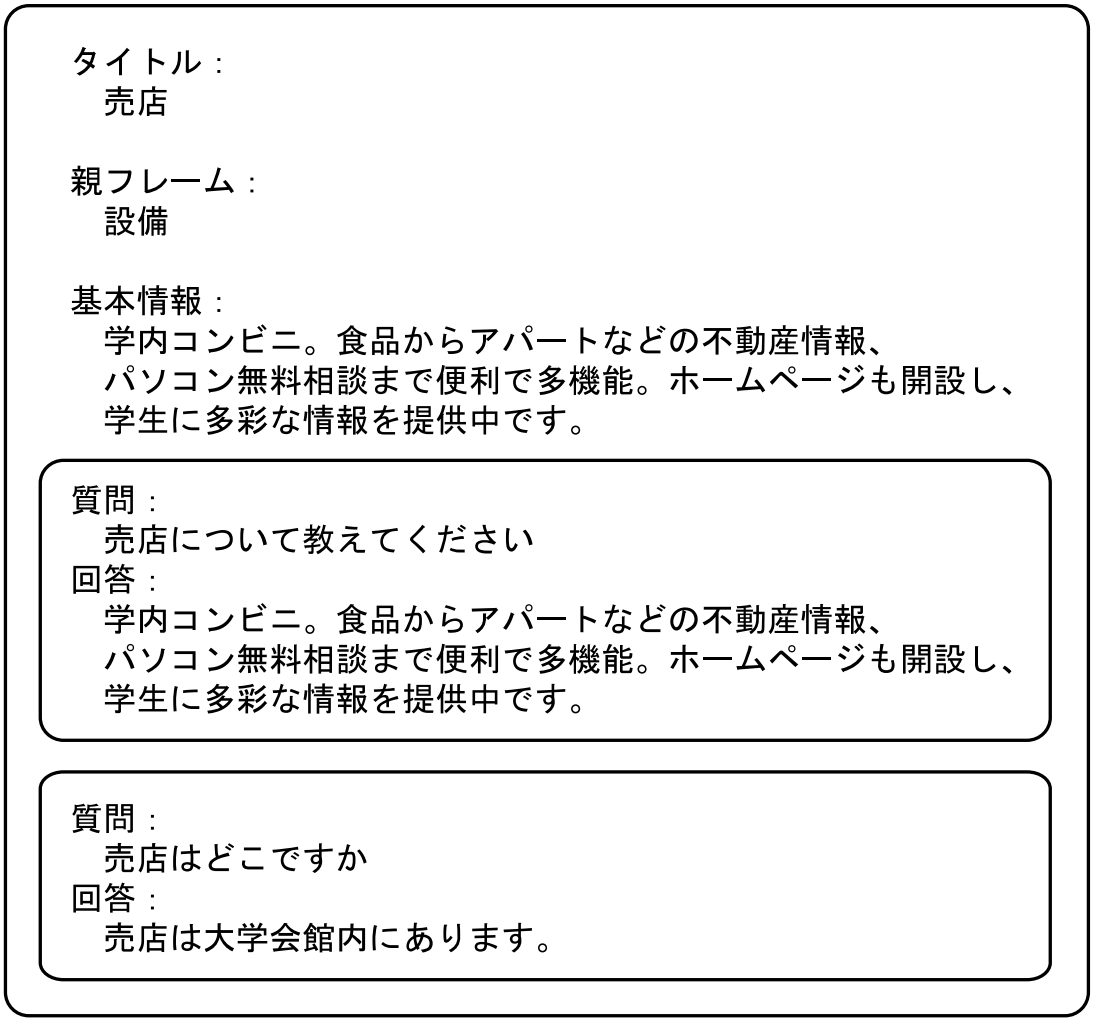


図 2.2　フレームの構成図

いる.

次に, 分解された形態素に重み付けがなされる. 図 2.5 は, 検索要求の質問のベクトル表現例を示す図である. たとえば, 「高知工科大学の特徴は何ですか」という検索要求の質問の中に含まれている索引語に「1」を, 含まれていない索引語に対して「0」を割り当てる. この操作により, 検索要求の質問に対して, 「1 1 1 1 0 0 0 0 0 0 1 0」というベクトルが作成される.

類似度の計算には, 内積を用いたベクトル空間法を利用する. 内積の計算は, 具体的に次のように行う. 検索要求の質問に含まれている索引語数 (ベクトルの次元) を m とし, 知識

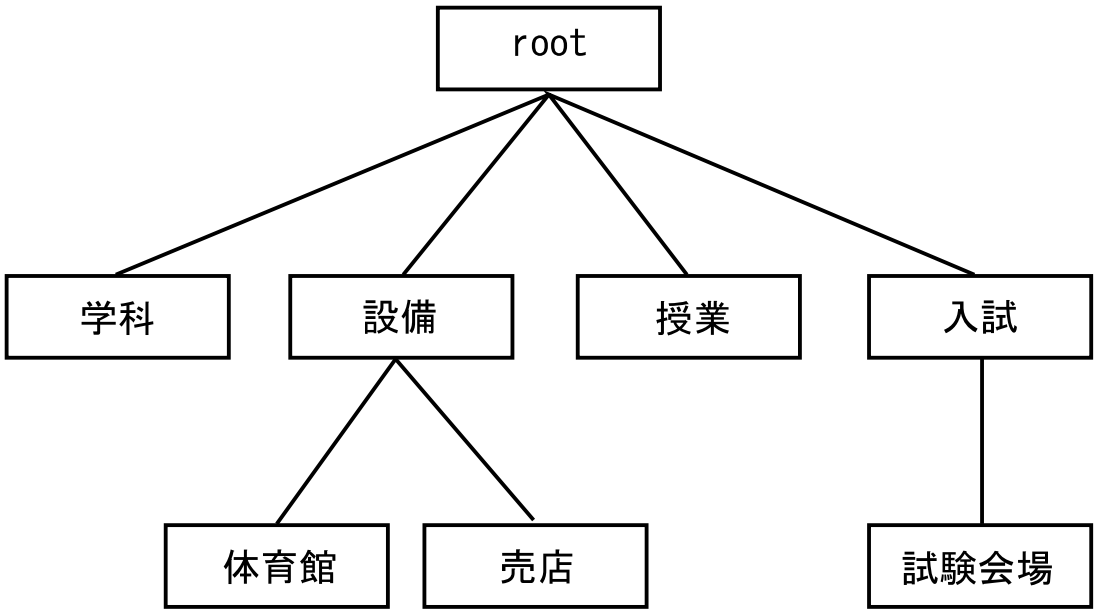


図 2.3　知識ベースの一部の構成図

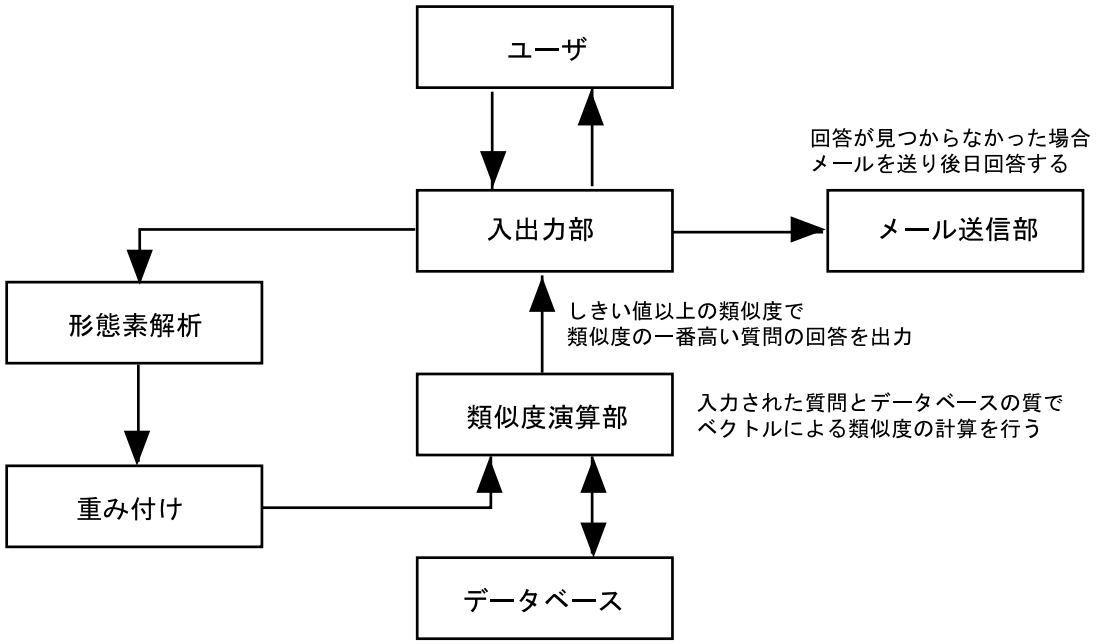


図 2.4　自動応答システムの構成図

検索要求の質問＼索引語	高知工科大学	の	特徴	何ですか	入試	試験科目	に	ついて	教えてください	学食	は	ありますか
高知工科大学の特徴は何ですか	1	1	1	1	0	0	0	0	0	0	1	0

図 2.5　検索要求の質問のベクトル表現例

ベース内の質問のベクトル x および検索要求の質問のベクトル y を

$$x = (x_1, x_2, ..., x_m) \tag{2.1}$$

$$y = (y_1, y_2, ..., y_m) \tag{2.2}$$

で表すとする. ベクトル x とベクトル y の類似度 $sim(x, y)$ は

$$sim(x, y) = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i^2}} \tag{2.3}$$

で表される.

式 2.3 により内積値を求め, 次 2.4 を満足するかどうかを判定する.

$$\text{(ベクトル化された検索要求の質問の内積)} >= \text{(しきい値)} \tag{2.4}$$

結果として, 式 2.4 を満たす知識ベースの質問とそのペアである回答が出力される.

2.4　質問応答

マッチングの結果により, 検索要求の質問に対する回答の方法は次のように変化する.

- 回答を返す

マッチングの結果, 式 2.4 を満たす場合, 類似度が最も高いデータを回答として出力する.

- 基本情報を返す

  マッチングの結果, 式 2.4 を満たさない場合, 基本情報が存在する場合は基本情報を回答として出力する.

- 回答を返さない

  上記以外の場合, 誤った回答を防ぐために, あえて多少類似したデータがあったとしても回答が見つからない旨を回答する.

第 3 章

ログ解析システムの構築

自動応答システムをはじめとする情報検索システムにおいて, 検索結果のログについて, 妥当性, 有効性などを解析することは, 情報検索システムの処理性能や検索に関する質の向上にとって極めて重要である.

このために, 従来は検索結果のすべてのログを保守者の目視により調査・評価する方法が採られてきた. 保守者は, 検索結果のログに記載された内容を参照することにより解析を行う. 従来のデータ解析支援プログラムでは, 保守者による解析処理を, GUI を工夫することによって支援している. しかし, 解析の対象となるすべてのログを 1 つ 1 つ保守者が解析しなくてはならない.

この方法では, 保守者への負担が大きい, 人為的なミスを排除できない, 保守者の判断が曖昧であり評価に一貫性が無いなどの欠点がある.

そこで, 本研究では, 保守者による目視の評価を事例データとしてデータベース化し, その後の検索結果のログを自動的に処理するインテリジェントなログ解析システムの開発に取り組んだ.

3.1 従来のデータ解析プログラム

従来のデータ解析およびデータ解析に対する支援プログラムについて説明する.

従来の情報検索システムの保守者は, 検索結果のログに記載された内容を参照することにより, 情報検索システムにおける検索結果のログの解析を行う. 従来のデータ解析支援プログラムでは, 保守者によるログデータを使用した解析処理を, GUI を工夫することによって

図 3.1　従来のデータ解析方法におけるログの一覧

支援している.

図 3.1 に, 従来の情報検索システムにおけるデータ解析支援プログラムの画面の例を示す. 図 3.1 に示す画面には, 受付番号, 利用者の検索要求の質問, ログの詳細な情報を表示するためのボタン, ログを解析するためのボタンが一覧形式で表示されている.

保守者がログの解析を行う場合, 保守者はログを解析するためのボタンを押し, 図 3.2 に示すログ解析画面を表示させる. 解析画面には, 検索要求の質問およびその質問に対して回答された内容が表示される. 保守者は, これらの情報から利用者とシステムの間でどのようなやりとりがあったのかを判断する.

保守者は, 質問と回答からなる検索結果のログに対してこのような解析を進めていく. このように, 従来の情報検索システムにおける検索結果のログの解析は, 解析の対象となる全データの一つ一つを人間が解析する.

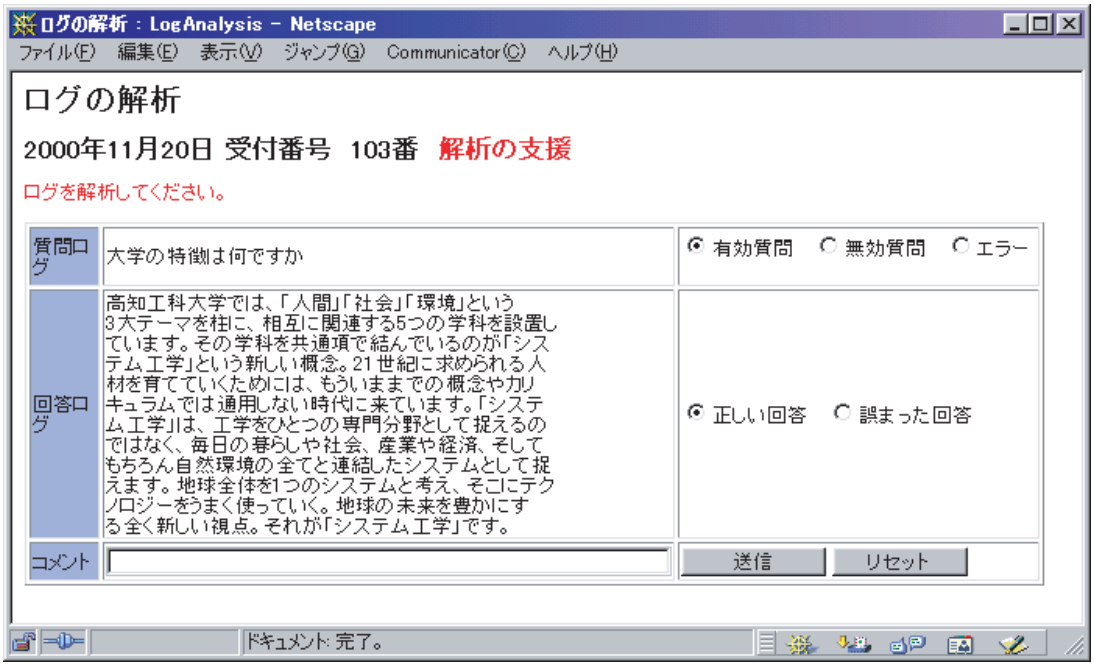


図 3.2　従来のデータ解析方法における解析画面

3.2　提案するログ解析システム

本論文で提案するログ解析システムについて説明する．本システムは，保守者による目視の評価を事例データとしてデータベース化し，その後に発生した検索結果のログを自動的に処理するインテリジェントなものである．

図 3.3 に，自動ログ解析のフローチャートを示す．

3.2.1　自動ログ解析

図 3.4 に，本ログ解析システムの自動解析の概要を示す．

事例データとは，過去の事例に関するデータであり，過去に保守者が解析を行った質問と回答およびその解析結果が納められている．事例データは，解析対象データを自動解析する際に使用される．また，事例データの集合体で事例データベースを構築している．

解析対象データは，自動応答システムが検索結果として回答した，利用者の検索要求の質問とその回答および中身が空の検索結果から成る．

3.2　提案するログ解析システム

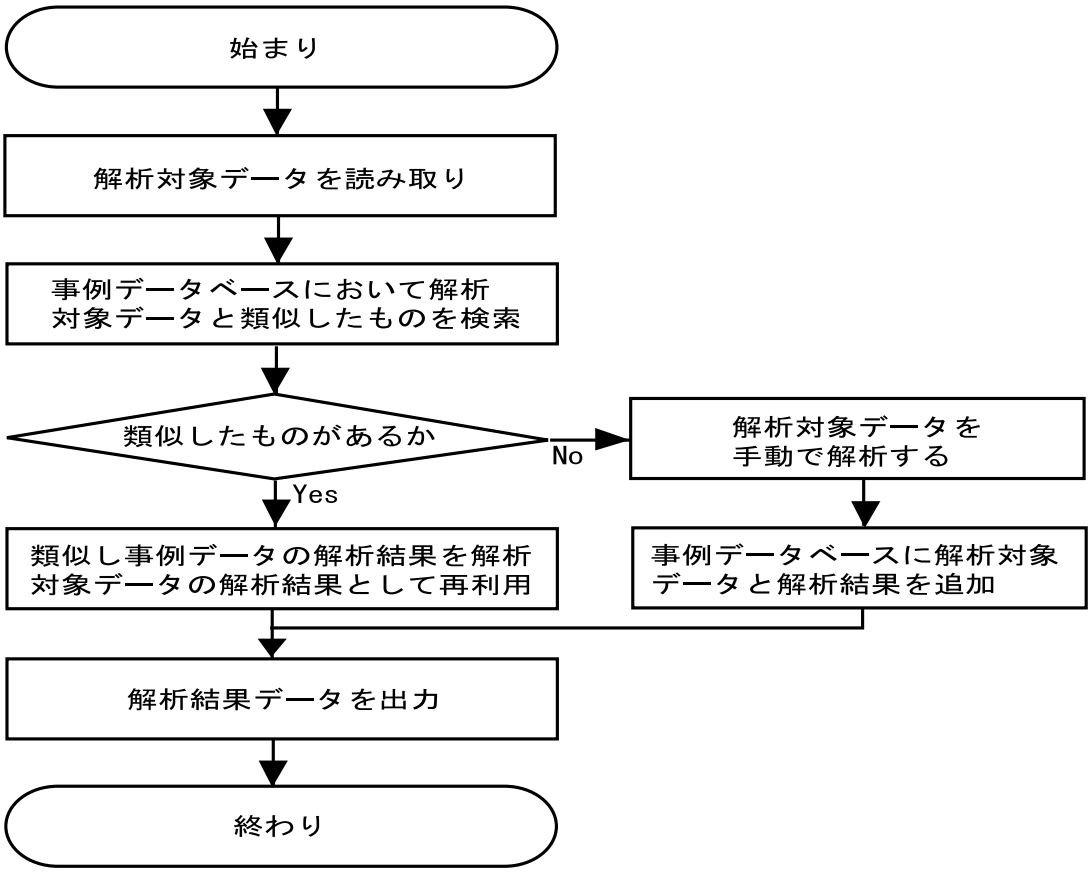


図 3.3　自動ログ解析のフローチャート

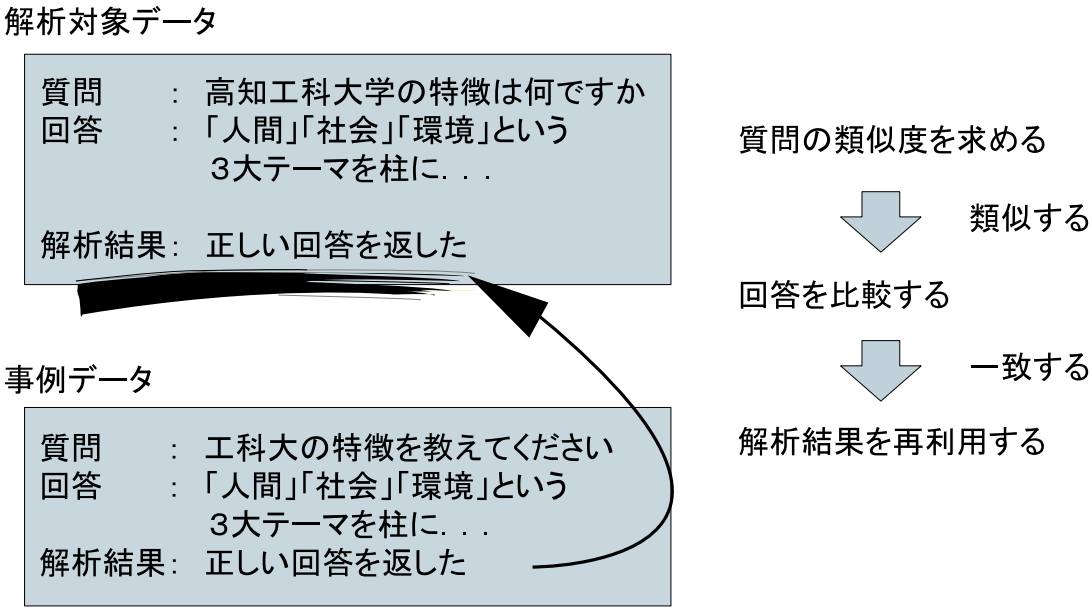


図 3.4　自動ログ解析の概要

索引語 / 検索要求の質問	高知工科大学	の	特徴	何ですか	入試	試験科目	に	ついて	教えてください	学食	は	ありますか
高知工科大学の特徴は何ですか	1	1	1	1	0	0	0	0	0	0	1	0

図 3.5　解析対象データの質問のベクトル表現例

まず, 解析対象データの質問と事例データの質問は, 形態素解析されて形態素に分解される. 形態素解析のツールとして, 計算機による日本語の解析の研究を目指す多くの研究者に共通に使える形態素解析ツールを提供するために開発された 茶筌 [4] を使用している.

次に, 分解された形態素に重み付けがなされる. 図 3.5 は, 検索要求の質問のベクトル表現例を示す図である. たとえば, 「高知工科大学の特徴は何ですか」という検索要求の質問の中に含まれている索引語に「1」を, 含まれていない索引語に対して「0」を割り当てる. この操作により, この質問による検索要求の質問に対して, 「1 1 1 1 0 0 0 0 0 0 1 0」というベクトルが作成される.

類似度の計算には, 内積を用いたベクトル空間法を利用する. 内積の計算は, 具体的に次のように行う. 解析対象データの質問に含まれている索引語数 (ベクトルの次元) を m とし, 事例データの質問のベクトル x および解析対象データの質問のベクトル y を

$$x = (x_1, x_2, ..., x_m) \tag{3.1}$$

$$y = (y_1, y_2, ..., y_m) \tag{3.2}$$

で表すとする. ベクトル x とベクトル y の類似度 $sim(x, y)$ は

$$sim(x, y) = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i^2}} \tag{3.3}$$

で表される.

式 3.3 により内積値を求め, 次 3.4 を満足するかどうかを判定する.

$$(\text{ベクトル化された解析対象データの質問の内積}) >= (\text{しきい値}) \tag{3.4}$$

結果として, 式 3.4 を満たす場合, 解析対象データの質問と事例データの質問とが類似しているという.

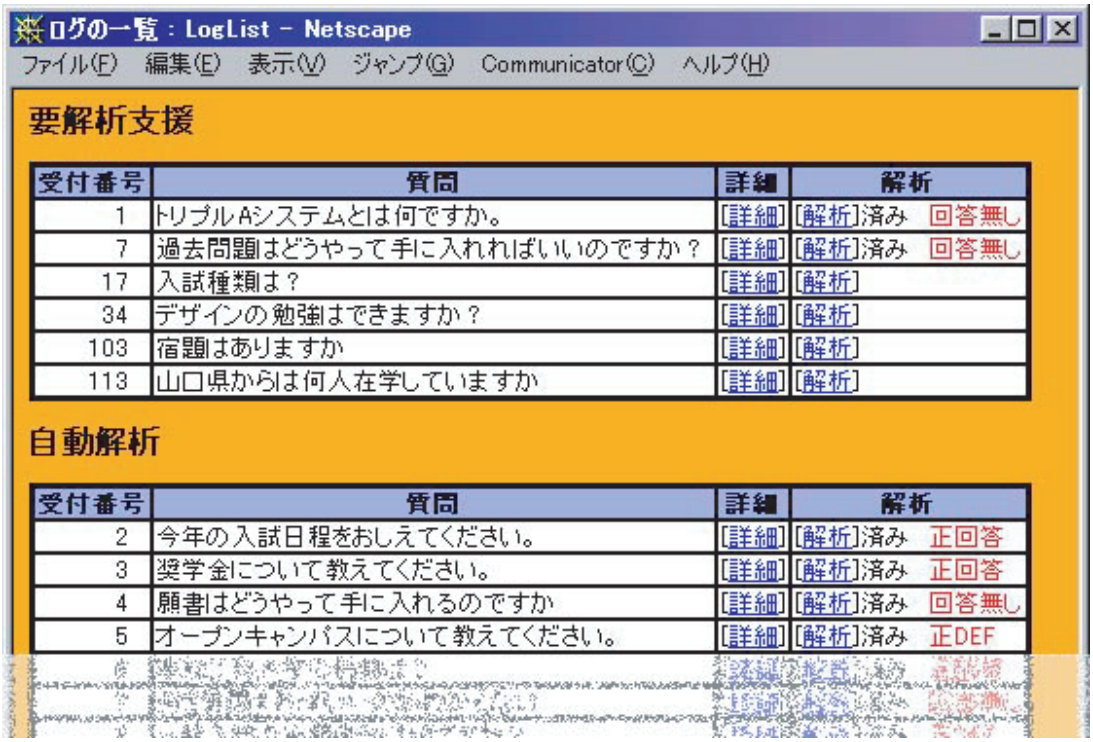
式 3.4 で類似していると判断されたとき, 次に, 解析対象データの回答と類似事例データの回答を比較する. ここで, 両回答が一致する場合は, 解析対象データと事例データは類似していると判断する. そして, 解析対象データと事例データが類似していると判断された場合, 解析対象データの解析結果として類似事例データの解析結果が再利用される.

図 3.6 に, 提案するログ解析システムの画面の例を示す. 図 3.6 に示す画面には, 受付番号, 利用者の検索要求の質問, ログの詳細な情報を表示するためのボタン, ログを解析するためのボタンが一覧形式で表示されている. 図 3.6 に示す画面には, 表が二つある. 下の表は, これまでに説明した方法で自動解析されたログの一覧である. ログを解析するボタンを押したときに表示される画面の例を図 3.7 に示す. 自動で解析された結果を見ることができる. 保守者は, 自動解析できなかった上の表にあるログだけを手動で解析すればよいのである.

3.2.2　解析支援

自動ログ解析を行った結果, 類似した事例データが無い場合は, 従来方式と同様に保守者は解析対象データの解析処理を支援する (図 3.2). そして, 解析結果が加わった解析対象データは, 事例データとして事例データベースへ追加され, 今後の自動ログ解析の際に利用される.

3.2　提案するログ解析システム


ログの一覧：LogList - Netscape

ファイル(F)　編集(E)　表示(V)　ジャンプ(G)　Communicator(C)　ヘルプ(H)

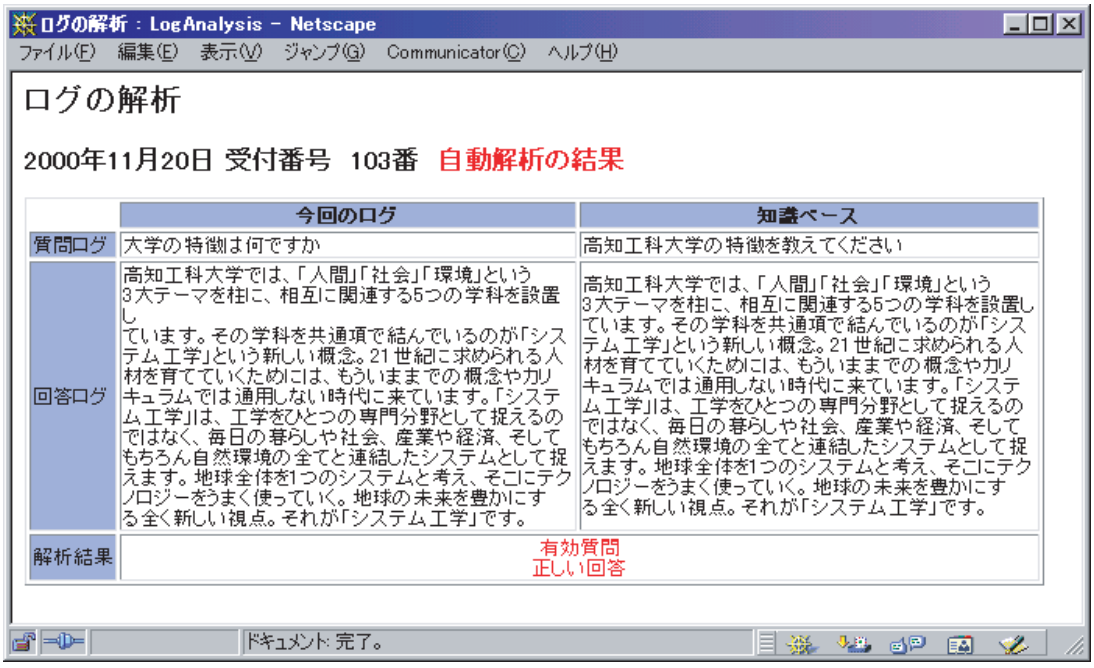
要解析支援

受付番号	質問	詳細	解析
1	トリプルAシステムとは何ですか。	[詳細]	[解析]済み　回答無し
7	過去問題はどうやって手に入れればいいのですか？	[詳細]	[解析]済み　回答無し
17	入試種類は？	[詳細]	[解析]
34	デザインの勉強はできますか？	[詳細]	[解析]
103	宿題はありますか	[詳細]	[解析]
113	山口県からは何人在学していますか	[詳細]	[解析]

自動解析

受付番号	質問	詳細	解析
2	今年の入試日程をおしえてください。	[詳細]	[解析]済み　正回答
3	奨学金について教えてください。	[詳細]	[解析]済み　正回答
4	願書はどうやって手に入れるのですか	[詳細]	[解析]済み　回答無し
5	オープンキャンパスについて教えてください。	[詳細]	[解析]済み　正DEF

図 3.6　提案するログ解析システムにおけるログの一覧


ログの解析：LogAnalysis - Netscape

ファイル(F)　編集(E)　表示(V)　ジャンプ(G)　Communicator(C)　ヘルプ(H)

ログの解析

2000年11月20日 受付番号　103番　自動解析の結果

	今回のログ	知識ベース
質問ログ	大学の特徴は何ですか	高知工科大学の特徴を教えてください
回答ログ	高知工科大学では、「人間」「社会」「環境」という3大テーマを柱に、相互に関連する5つの学科を設置し ています。その学科を共通項で結んでいるのが「システム工学」という新しい概念。21世紀に求められる人材を育てていくためには、もういままでの概念やカリキュラムでは通用しない時代に来ています。「システム工学」は、工学をひとつの専門分野として捉えるのではなく、毎日の暮らしや社会、産業や経済、そしてもちろん自然環境の全てと連結したシステムとして捉えます。地球全体を1つのシステムと考え、そこにテクノロジーをうまく使っていく。地球の未来を豊かにする全く新しい視点。それが「システム工学」です。	高知工科大学では、「人間」「社会」「環境」という3大テーマを柱に、相互に関連する5つの学科を設置しています。その学科を共通項で結んでいるのが「システム工学」という新しい概念。21世紀に求められる人材を育てていくためには、もういままでの概念やカリキュラムでは通用しない時代に来ています。「システム工学」は、工学をひとつの専門分野として捉えるのではなく、毎日の暮らしや社会、産業や経済、そしてもちろん自然環境の全てと連結したシステムとして捉えます。地球全体を1つのシステムと考え、そこにテクノロジーをうまく使っていく。地球の未来を豊かにする全く新しい視点。それが「システム工学」です。
解析結果	有効質問 正しい回答	

ドキュメント: 完了。

図 3.7　提案するログ解析システムにおける自動解析結果画面

第 4 章

解析データの評価

高知工科大学ヘルプシステムは, 2000 年 9 月 15 日より本学のホームページで公開し試験運用している. 公開から 11 週間経過する間に受け付けた質問は約 1800 件であり, 検索結果のログはすべて整理し保存されている.

本章では, 約 1800 の検索結果のログを本論文で提案するログ解析システムを用いて解析を行う. また, 現時点での高知工科大学ヘルプシステムの検索結果における失敗例から, 今後の同ヘルプシステムに必要な知識や機能について考察を行う.

4.1 ログ解析システムの評価

本論文で提案するログ解析システムの性能を評価する実験を行った.

本来, 本システムは, 保守者の手動でのログ解析結果を蓄積した事例データベースを構築したうえで利用されることを想定している. 実験では, 事例データがまったく無い状態からどれくらいのログが自動で解析できるかを測定した. また, 誤解析の可能性を低減するために, 類似度の計算の際に使用するしきい値は最大まで厳しくした.

この結果, 約 1800 件のログにおいて約 13.6%については自動的に解析を行うことが確認できた. 事例データの数が多くなり, 適切なしきい値を与えたならば, さらに自動解析を行うことができると期待できる.

4.2 解析したログの評価

まず, 利用者からの質問を 3 つのタイプに分類した.

正しい回答を返した割合	33.2%
誤った回答を返した割合	1.7%
正しい基本情報を返した割合	31.2%
誤った基本情報を返した割合	4.9%
回答を返さなかった割合	29.0%

表 4.1　高知工科大学ヘルプシステムの解析結果

- 有効質問

  自動応答システムが想定している知識ドメインに対する質問.

- 無効質問

  自動応答システムが想定している知識ドメインを極端に逸脱している質問. または, 質問と受け取ることができない文章や文字列.

- エラー

  有効質問とも無効質問とも取ることができない質問や文章や文字列. およそ, 文字化けなどの原因によるものが多い.

これまでの全期間の有効質問における成績は以下の通りである.

4.3　利用者の質問の特徴

利用者から寄せられる検索要求の質問には, ある程度特徴があることがわかった. 以下に, その特徴をまとめる.

4.3.1　単語のみの質問

本論文で述べる自動応答システムでは, 従来のキーワード型の情報検索システムと違い, 自然言語で書かれた質問に対して回答を返すことが可能であることを特徴としている.

しかし, 利用者は, このキーワード型の情報検索システムと同様に考え単語で質問したと

質問
入試日程
ドミトリー
授業料
学内ネットワーク
ＬＡＮ

表 4.2　単語のみの質問の例

も考えられる. この事例は多数見られるので, 単語のみの質問に対しても, 自然言語で書かれた質問と同等に回答できる機能が必要と考えられる.

表 4.2 に, 単語のみの質問の例を示す.

4.3.2　語尾が無い質問

本論文で述べる自動応答システムは, 質問パターンを登録しておくことで言い回しによる質問の違いを理解している. これまで, 語尾が無い質問の場合, 経験則で What is 型で処理を行ってきた.

ログ解析の結果語尾の無い質問については, What is 型の質問パターンで処理できることがわかった.

表 4.3 に, 語尾が無い質問の例を示す.

4.3.3　複数の意味がある質問

通常, 一つの質問には一つの回答を返すものである. しかし, 一つの質問文中に二つ以上の回答を返す必要がある場合があることがわかった.

たとえば, 「○○と△△について教えてください」という質問の場合, ○○についてと△△についての二つの質問について回答する必要がある.

質問
高知工科大学の特徴
願書の請求方法
水の再利用
今日は何曜日
学食の場所

表 4.3　語尾が無い質問の例

質問
受験料、入学金はいくらですか？
入学金、学費、について知りたい
学校の特色や校訓を教えてください。
受験科目はなんですか、難しいですか。
取得できる資格と、就職状況を教えてください

表 4.4　複数の意味がある質問の例

表 4.4 に, 複数の意味のある質問の例を示す.

4.3.4　あいさつ

本論文で述べる自動応答システムは, 自然言語で書かれた質問に対して適切な回答を返す特徴を持つ. より人間に近いコンピュータシステムを実現していることから, 利用者がまるで人間に話しかけるようにあいさつを入力したものと思われる. しかし, 現在の本システムにあいさつなどの会話を行う機能は無い.

表 4.5 に, あいさつの例を示す.

質問
こんばんわ
こんにちは

表 4.5　あいさつの例

質問
すごいんですね
まともに答える気がありますか
このヘルプシステムは、とても興味深くて面白かったです。
いいんじゃない？
あなた使えませんね

表 4.6　利用者の本自動応答システムに対する評価の例

4.3.5　利用者の本自動応答システムに対する評価

本論文で述べる自動応答システムには, 利用者が本システムを評価できる機能が無い. このことから利用者は, 質問として本システムの評価をしたものと思われる.

表 4.6 に, 複数の意味のある質問の例を示す.

4.3.6　言葉使いを一部変えて続けて同じ内容の質問

利用者がシステムの回答に満足できず, 言葉使いを一部変えて質問し直すときに, 本論文で述べる自動応答システムは再び同じ回答を返してしまう場合がある. 意味的まとまりを持つ連続した質問をセッションと呼ぶと, 本システムにはセッションを理解する機能が無い. したがって本システムは, 続けて似たような質問を受け付けても, これらの質問を同セッション中の質問として理解することはできない.

表 4.7 に, 同じ回答を返した質問の対の例を示す.

質問
例 1 ）高知工科大学はいつできたのですか
高知工科大学は何年にできたのですか
例 2 ）工科大の URL はなんですか？
工科大の URL を教えて下さい。
例 3 ）高知工科大学にある学部は？
高知工科大学には何学部があるのですか。

表 4.7　言葉使いを一部変えて続けて同じ内容の質問の例

質問
1 回目）高知工科大の寮はありますか
2 回目）何部屋ありますか

表 4.8　前回の続きの質問の例

4.3.7　前回の続きの質問

表 4.8 をみると, 利用者がシステムの回答を受け, 続けてさらに詳しい質問をするときに, 主語を省略して質問をしたことがわかる.

本論文で述べる自動応答システムはセッションを管理する機能が無いため, 主語が抜けた質問に対して回答することができなかった.

4.3.8　知識ドメイン外の質問

高知工科大学ヘルプシステムでは, 本学への入学を目指す受験生を対象とし, おもに受験に対する疑問・質問に答えるための知識ベースを構築している. このため, 表 4.9 に示すような質問は, 同ヘルプシステムにとって知識ドメイン外の質問となる.

しかし, 本論文で述べる自動応答システムは, 知識ベースを自然言語を用いて構築するこ

質問
高知県にある放送局を教えてください
日本の人口は
四国は何県ありますか
日本で一番高い山は
日本の総理大臣は

表 4.9　知識ドメイン外の質問の例

質問
高知行きのバスの便は
高知工科大のキャンパスりの
受験システムってどうなってるの？

表 4.10　半角カナが混じっている質問の例

とができる. したがって, 比較的容易に本ヘルプシステムのような本自動応答システムの応用システムが構築できる.

4.3.9　半角カナが混じっている質問

本論文で述べる自動応答システムでは, 半角カナ文字及び半角英数文字はシステム内部では全角文字へと変換されて処理される. これにより検索効率を高めることを可能としている.

ところが, 表 4.10 を見ると濁点や半濁点が含まれる場合には注意が必要であることがわかる. 半角のバスを全角に変換すると「ﾊﾞｽ」となり「バス」ではなく, 同様にキャンパスの場合は「キャンﾊﾟｽ」であり「キャンパス」ではない.

表 4.10 の例では半角カナを表現できないため全角で表記する. いずれもカタカナが本来半角の部分である.

質問
親戚の者が工科大学を推薦で受験します。(情報システム工学科を専願)
親戚の者が情報システム工学科を推薦入試で受験します。(専願)
このせっぱ詰まったときにはどのような勉強をしたらよいのかと悩んでいます。
教えてください。

表 4.11　かっこ書きを含む質問の例

4.3.10　かっこ書きを含む質問

本論文で述べる自動応答システムは, あたかも人間同士が会話をするかのように質問することができる. 表 4.11 の例を見ると, 人間同士のやりとりには少なくとも二つのパターンがあることがわかった. つまり, 書き言葉と話し言葉である. 話し言葉であれば, 例のようなかっこ書きを用いた表現は行わない. しかし, 書き言葉にはこれがある. かっこの中身が意味的に重要である場合があるので, かっこ書きを含む質問には特別の処理を施す必要があるかも知れない.

4.3.11　構文情報を必要とする質問

本論文で述べる自動応答システムは, 構文解析を行わない. したがって, 本システムは構文の意味的に質問のどの言葉が大切であるかを理解することができない.

このために誤った回答になった質問の例を表 4.12 に示す.

4.3.12　綴り間違いがある質問

人間は, たとえ間違った文章または文字が書かれていても, ある程度解釈して読み換えることができる. しかし, 本論文で述べる自動応答システムにはこのような機能はない.

システムにとって曖昧な言葉や理解できない言葉は, 聞き返しを行うなどの対処法を行うべきだと考えられる.

質問
高知工科大学の図書館の特徴を教えてください。
情報システム工学科の特徴を教えてください
学生の駐車場はありますか？
高知工科大学の住所の英文スペルを教えて下さい。
山田駅からのアクセスは

表 4.12　構文情報を必要とする質問の例

質問
学際はありますか
受験科は
ドミトリーについて教えてください
高知工科大学に s h
j 競争率は？

表 4.13　綴り間違いがある質問の例

表 4.13 に, 綴り間違いがある質問の例を示す.

4.3.13　UNIX コマンドなど

これらの質問は UNIX コマンドもしくはそれに関連する言葉である. 本論文で述べる自動応答システムは, 半角英数文字をシステムの内部では全角文字へと変換されて処理するため, 本システムがこれらを UNIX コマンドとして理解して誤作動を起こすことはない.

表 4.14 に, UNIX コマンドなどの例を示す.

質問
telnet
ftp
rlogin
gnutella
SMTP

表 4.14　UNIX コマンドなどの例

4.4　知識ベースに不足しているフレーム

知識ベースに蓄えられているデータが不足しているために, 検索要求の質問に対して回答ができないことがあった. おもに以下に示す分野について多くの場合に回答することができなった.

- 「教授」,「助教授」などの教員に関する質問
- 「学生数」,「男女比」などの統計データに関する質問
- 「学部」に関する質問
- 「試験日」に関する質問
- 「シラバス」,「教育方針」に関する質問
- 「募集要項」に関する質問
- 「試験会場」に関する質問
- 「研究」に関する質問
- 「開学」に関する質問
- 「連絡先」に関する質問

第 5 章

おわりに

本論文では, 事例データを活用して自動的に解析を行うインテリジェントなログ解析システムを提案した. 本システムは実際に現在試験運用中の高知工科大学ヘルプシステムのログ解析システムとして日常的に利用している.

本システムを利用して解析を行った結果, 高知工科大学ヘルプシステムのこれまでの期間の回答率は, 正しい回答の 33.2%と正しい基本情報の 31.2%の合計で 64.4%であったことがわかった. また, 利用者の質問から特徴を見ることができ, 同ヘルプシステムの性能向上へのあしがかりとなった.

本研究により, 検索結果のログ解析において保守者の負担を低減し, これらの解析結果や, 自動応答システムの問題点を見つけることができた. 保守者の負担が低減できたことにより, 人為的なミスによる誤解析の確率を少なくし, 一貫性のある結果を得ることが可能となった.

今後の課題として, 自動応答システムの知識ベースにおいて, 知識が不足しているフレームを自動的に割り出すことが上げられる. これを実現することで, 検索結果のログを自動的に解析するばかりでなく, 自動的に評価することが可能となり, 保守者の負担がさらに低減するものと考えられる. また, 解析速度の向上も課題としてあげられる. 本研究の現在のログ解析システムでは, 自動解析の際に無駄が多々あり解析速度の低下を招いている. 無駄を省く, または, 新しい自動解析のアルゴリズムを研究することも今後の課題となった.

謝辞

本研究を進めるにあたり終始懇切丁寧に御指導下さいました Ruck Thawonmas 助教授に心より御礼申し上げます.

また, 本研究のログ解析システムへのアドバイスのみならず, UNIX についても御教授いただいた, 株式会社 エス・エス・アールの溝渕 真司氏ならびに吉岡 倍達社長をはじめ社員の方々にも心から御礼申し上げます.

高知工科大学ヘルプシステムの試験運用を快く承知いただき, 知識ベースの構築に御協力いただいた, 教務学生課 入試班ならびに関係者各位にも, 心からの謝意を述べさせていただきます.

大学生活を送るにあたり, 研究活動のみならず私生活でもお世話になった親友たち, ありがとう.

最後に, 本研究に関して援助して下さったラック研究室の皆様, 特に本論文の締め切りが近づくとマンションが同じなのをいいことに連日車通学に便乗させていただいた 平山 純一郎氏に, 本当にありがとうございました.

参考文献

[1] “平成12年度版通信白書”, http://www.mpt.go.jp/policyreports/japanese/papers/h12/.

[2] 友池 貴之, 溝渕 真司, ラック ターウォンマット, “データ解析方法およびその装置ならびにコンピュータ・プログラム”, 特許出願中 (審査請求), Jan.30,2001.

[3] 長尾 真, “岩波講座 ソフトウェア科学 15 自然言語処理”, 岩波書店, pp.411-417, 1996.

[4] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸, “形態素解析システム『茶筌』version 2.2.1 使用説明書”, 奈良先端科学技術大学院大学 松本研究室, pp.1, 2000.