分散環境における人間の頭脳を利用したデータ収集システム

尾崎 亮太

電気通信大学 情報工学科 中山泰一研究室

E-mail: ozaki–r@igo.cs.uec.ac.jp

概 要

インターネットに蓄積されたデータは利用者の増加に伴い，多様かつ膨大なものになっている．その膨大なデータの中には大量の知識が含まれていることから，インターネットを巨大な知識ベースとみなすことが可能である．

知識ベースとしての能力を高めるためには，インターネット上にいる人間が持つ知識のデータ化を促進し，インターネット上の知識データを増加させることが重要である．

掲示板やメーリングリストなどでは，ある人の質問・疑問などに他の人々が回答することでコミュニケーションを行なっており，そのログが知識データとしてインターネットに蓄積されている．このように知識のデータ化には人間の質問・疑問などの情報が重要な役割を果たしていると考えられる．

本稿では，現在の知識のデータ化プロセスを考察しその問題点を挙げ，その問題点を踏まえたシステムを提案する．提案するシステムは SETI@home のような広域分散システムを採用し，質問・疑問を多くの人間に提示し，その情報を管理する機構を用意する．また知識データの妥当性を保障するための機構を導入する．

1 はじめに

Web サーフィンや電子メールの流行に伴いインターネット人口は増加の一途をたどっている．現在インターネット人口は 5 億人を越えており，世界の全人口の 10 分の 1 に達するのも時間の問題である．

インターネットは WWW という誰でも利用可能な情報発信の手段を手に入れたのち，その利用人口を劇的に増加させた．それまでは一部の技術者や学術関係者だけの利用に留まっていたインターネットだが，今や一般人の生活の一部となりつつある．

インターネットに蓄積されたデータは利用者の増加に伴い，多様かつ膨大なものになっている．その膨大なデータの中には大量の知識が含まれていることから，インターネットを巨大な知識ベースとみなすことが可能である．実際日常生活において何か疑問に思うことがあった場合，検索サイトを用いてその答えを探すといったことが行われている．またインターネットから知識を効率良く引き出す研究なども行われている [1].

しかしインターネットから取り出すことのできる知識はデータ化され蓄積された知識 — 本稿ではこれを知識データと呼ぶ — のみである．たとえインターネットに接続している人間が知識を保有していたとしても，その他の人間はその知識を取り出すことができない．インターネット上にいる人間が持つ知識のデータ化を促進し，インターネット上の知識データを増加させることが重要である．

掲示板やメーリングリストなどでは，ある人の質問・疑問などに他の人々が回答することでコミュニケーションを行なっており，そのログが知識データとしてインターネットに蓄積されている．このように知識のデータ化には人間の質問・疑問などの情報が重要な役割を果たしていると考えられる．

本研究では質問・疑問の重要性に着目し，その情報を利用することで知識のデータ化を支援するシステムを提案する．

本稿では，現在の知識のデータ化プロセスを考察しその問題点を挙げ，その問題点を踏まえたシステムを提案する．提案するシステムは SETI@home [3] のような広域分散システムを採用し，質問・疑問を多くの人間に提示し，その情報を管理する機構を用意する．また知識データの妥当性を保障するための機構を導入する．

2 背景

2.1 知識データ

インターネット上にある知識データは，以下のような形で蓄積されていると考えられる．

- 通常の情報発信を目的として書かれた Web ページ
- 掲示板やメーリングリストなどのログ

前者は，個人もしくは組織の持つ知識をそれぞれの理由でデータ化する場合である．個人レベルでは自分の知っている情報をまとめて自身の Web ページで公開することが行われている．最近は日記を Web ページで公開する Web 日記が流行っており，その中に自分が手に入れた有用な情報などを書き残すことが行われている．この個人の日課が結果として知識データとなりインターネットに蓄えられる．また企業などでは自社のサービスの FAQ や技術情報を Web ページとして公開しており，これもまた知識データに含まれる．

後者の場合は元々なんらかの情報のやり取りが主目的でありその結果生成されたログが知識データとなる．掲示板やメーリングリストなどではある人間が質問，話題を投稿しそれに対する回答をやりとりすることでコミュニケーションが行われる．この場合質問・回答の組が知識データとなる．

以上の状況から二つのことが判る．一つは，インターネットには Web ページを書いたり，他人の質問や疑問に答えるという事を積極的にする人が多いということ．もう一つは，何らかの目的やきっかけが知識のデータ化を促進させ，そのきっかけとして質問・疑問が重要な役割を果たしている，ということである．

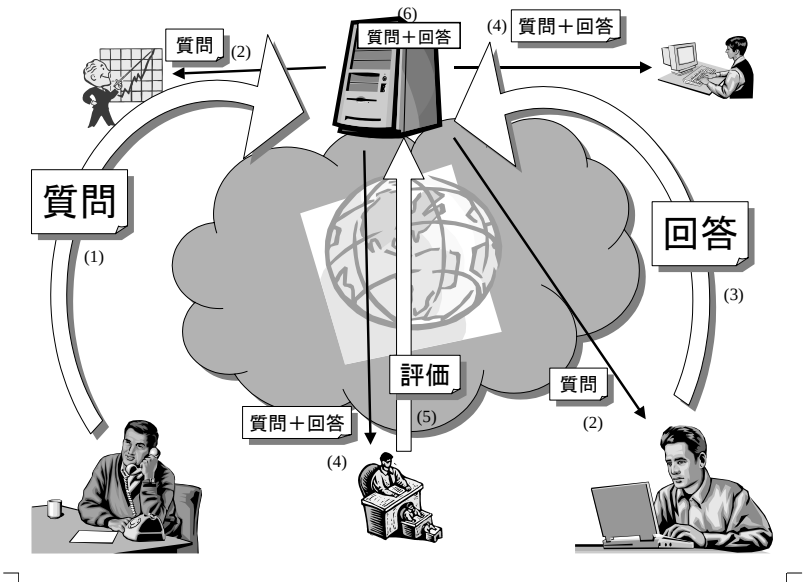
2.2 問題点

掲示板やメーリングリストなどのコミュニケーションシステムはインターネット上で，ある特定の人間が参加しお互いに情報交換を行う手段として最も一般的であると言える．そのコミュニティは互いに興味や趣味の近い人間の集まりであることが多く，コミュニケーションの場としては有効である．

しかし，このコミュニケーションシステムを知識をデータ化する手段として考えると以下のような問題点が存在する．

- 特定の人間しか見ない
  掲示板システムの場合はその掲示板を頻繁に見に行く人，メーリングリストであるならば参加している人をそのコミュニティに属しているとみなすことができる．質問・疑問は基本的にコミュニティに属している人間しか知ることはない．
  インターネット上には非常に多くの人間がおり，潜在的に答えを知っている人間が存在する可能性は高い．質問・疑問がその人間に知られないことは，知識のデータ化の機会を失うという意味で多大な損失といえる．
- 参加者でないと書き込めない
  会員制掲示板やメーリングリストの多くは，そのシステムへ登録を行わなければ利用することができない．参加が自由なシステムであっても，コミュニティとして閉じており，他者が容易には質問ができない状況も多い．
  このように，何らかの疑問を持った人間が質問したくてもコミュニティに属していないために，それが行えないことがありうる．この場合，質問・疑問のデータ化すら行われない可能性がある．
- 質問・疑問が喪失する可能性がある
  もしそのコミュニティ内にその答えを知っている人がいなかった場合，その質問・疑問は忘れさられてしまう．現存のコミュニティシステムは，構造的に古い情報は新しい情報に比べ閲覧し難くなるため，その質問・疑問を再度誰かが見る可能性は低い．もし回答を知っている人が見たとしても，回答することができない，またはでき難くなっていることが多い．(古くなった質問・疑問への回答はほとんど行われないのが現状である)
- 知識データが喪失する可能性がある
  掲示板システムの場合，生成された知識データが容量制限のために無くなってしまう可能性がある．失われた知識データは検索サイトのキャッシュに残っている可能性もあるが，いずれ失われる．
- 知識データが整理されていない
  元々知識データを蓄えることを目的としていないため，データが構造化されておらず，また余計な情報も多く含んでいる．
  メーリングリストや掲示板では，人々の書き込みが単に前の書き込みに付け足される形で保存される．そのため質問・疑問と回答が離れてしまい，非常に読みづらいことがある．またキーワードが離れてしまい，検索するときうまく見付からないこともある．

このような問題点により，既存のコミュニケーションシステムは知識データの蓄積を行う用途には不向きであると考えられる．インターネットの知識ベースとしての利便性を高めるためには新たな知識データ蓄積のメカニズムが必要である．そこで本研究では，インターネット上の知識データを効率良く増やすシステムを提案する．



(1) 質問をする．

(2) その質問が配布される．

(3) その質問に回答する．

(4) 回答に対する評価の仕事が配布される．

(5) 評価の仕事を行う．

(6) 知識データが蓄えられる．

図 1: システム概要

3 提案するシステム

前章の問題点を踏まえると，知識データを効率良く蓄積するためには以下のような要件を満たすシステムが必要であると考えられる．

- 多くの人々が質問・疑問を見ることができる．また多くの人々がそれに回答することができる．
- 質問・疑問および知識データを管理する機構が存在する．

3.1 システム概要

本研究では SETI@home や Grid システムのような広域分散システムの構造を採用する．インターネット上にシステムを分散させ，質問・疑問を広域に配布することで多くの人がそれを見る機会を増加させる (図 1)．

質問・疑問を投げかけたり回答を行うためのインタフェースは，SETI@home のように参加者に配布したクライアントプログラムであったり，Web ページの広告欄などである．参加者はそのインタフェースを通して質問や回答を行う．

本システムでは質問・疑問やその回答を管理する．質問・疑問の配布や，回答の回収，またそれらの蓄積などを行う．本システムでは質問・疑問と回答の組を知識データと考え，これを蓄える．蓄えられた知識データは，通常の Web ページの形式で公開されるため，通常の検索サイトで検索可能である．

以下でシステムの各要素を詳しく述べる．

3.2 広域分散システム

本システムにおいて質問・疑問を多くの人に見てもらうことは，最も重要な目標の一つである．既存のコミュニケーションシステムにおいては，参加者しか質問・疑問を見ることがない．参加者に答えを知っている人間がいないことがしばしばある．参加者以外の人間がその答えを知っている可能性は高く，その人間が質問・疑問を見ないことは非常に勿体ない．そこでシステムを分散させ潜在的回答者が質問・疑問を見る機会を増やす．

SETI@home では参加者に仕事として計算問題を配布する．本システムでは計算問題の代わりに質問・疑問などの仕事を配布する．参加者は質問・疑問の回答などを結果としてシステムに返す．

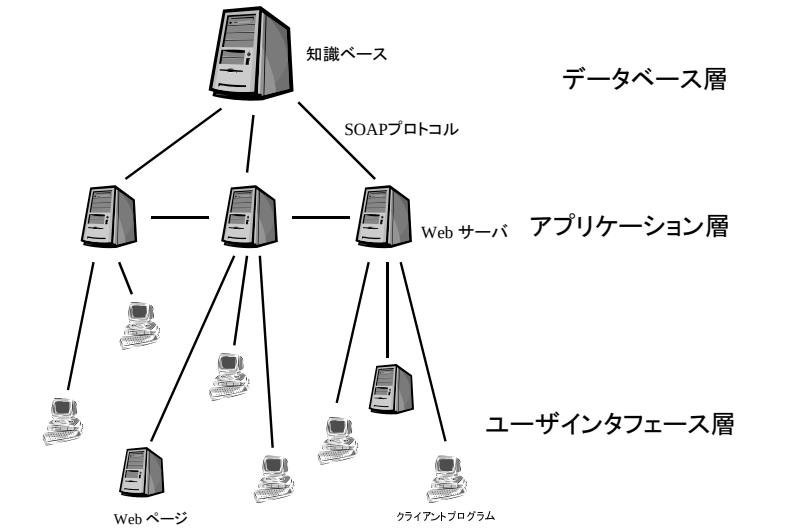


図 2: システム構成

本システムではシステム内の通信プロトコルに SOAP (Simple Object Access Protocol) [4] を使う．SOAP は分散環境における情報交換のためのプロトコルであり，他の HTTP や SMTP と組み合わせて使うことが可能である．そのため，インターネット上の様々なシステムをインタフェースに利用することが可能となる．例えば参加者の計算機で実行されているクライアントプログラムであったり，Web ページの広告などをインタフェースにすることが非常に容易である．また SOAP を使うと他のサービスとの連係が可能である．

本システムは図 2 のように 3 層モデルになっている．

集められた知識データはデータベース層で保存される．仕事の管理はアプリケーション層で行われる．質問・疑問を表示するインタフェースはユーザインタフェース層となっている．

アプリケーション層は通常 Web サーバであり，ユーザインタフェース層は Web ページの広告であったり，参加者がダウンロードしたクライアントプログラムである．Web サーバや Web ページの広告スペースを利用するには，有志を募る必要がある．また参加者は Web サーバやクライアントプログラムをシステムに登録する必要がある．

3.3 知識データ管理

まず本システムで扱う知識データについて述べる．知識データの形式は質問・疑問とその回答の組とする．この制限はシステムで知識データを管理しやすくすると同時に，知識データを検索しやすくするなどの効果がある．また質問・疑問は答えが一意に決まるようなもののみを扱う．

本システムでは質問・疑問の配布や回答の回収，またそれらの蓄積などを行う．それに加えシステムでは知識データの品質を管理する必要がある．

質問・疑問やその回答は必ずしも品質が良いとは限らない．どのような回答を求めているか判らない質問・疑問や，適切でない回答は切り捨てる必要である．

また複数回答に対する対処を考える必要がある．本システムにおいては質問・疑問が複数個配布されるため，当然回答が複数個返ってくる可能性がある．一つの質問・疑問に複数の回答が存在することは，データとしては冗長であり，またその知識データを見る人にとっては混乱の元となる．

そこで知識データの品質を保障するために，質問・疑問やその回答を評価する仕事と，複数の回答を比較する仕事を導入する．

- 妥当性評価
  仕事の結果として返ってきた回答に対して，その妥当性を問う仕事を新たに生成し配布する．参加者はその回答が妥当であるか判定し，その結果を返す．システムはこの仕事の結果を集計し，回答の正当性の判断を行う．妥当であると判断された回答は知識データの候補として残される．妥当でないと判断された回答は候補から外される．
- 比較
  同じ質問・疑問へ複数の回答が返された場合，その中のうちより良い回答はどれであるか判断を行う．参加者には複数の回答が提示される．参加者はその回答の中から優れているものを選択する．より優れていると判断されたものが残り，そうでないものは削除される．

この仕事は二者択一問題であり，非常に簡単な仕事である．このような簡単な仕事であれば，怠惰な人間であっても仕事をしてくれる可能性が高い．

3.4 問題点に対する解

ここでは，全節における問題点を提案するシステムでどのように克服するかについて述べる．

- 特定の人間しか参加しない・参加者でないと書き込めない
  広域分散システムを採用し，どこからでも誰でも本システムに対してアクセス可能である．そのため質問・疑問に対する回答が得られる可能性を高めることができる．
- 質問・疑問が喪失する可能性がある・知識データが喪失する可能性がある
  質問・疑問や回答などのデータはシステムで管理する．そのため，既存のコミュニケーションシステムのようなデータの喪失は起きない．例え回答がなかなか返ってこない質問・疑問であっても，データは保持され続ける．
  しかし妥当性がないと判断された回答などの情報は削除される．
- 知識データが整理されていない
  本システムでは，知識データを質問・疑問とその回答の組の形で扱う．この制限により知識データは常に整理された状態を維持できる．また不要な回答は評価機構によって取り除かれ，基本的に単一の回答しか残らないため可読性も向上する．

4 課題と解決法

ここでは前章で提案したシステムにおいて，解決しなければならない問題点を挙げ，その解決法を述べる．

4.1 仕事の配布先

仕事を広域に配布することは，潜在的回答者が質問・疑問を見る可能性を高める．しかし，必要とする知識が異なる質問・疑問をだれかれ無しに配布することは効率的でない．また参加者は得意分野でない質問・疑問の答えをばかり問われたくはないであろう．

この問題を解決するために，本システムでは質問・疑問の分類を行う．Web サーバやクライアントプログラムを登録するとき，自分の得意分野を指定し配布される仕事を制限する．

質問・疑問を分類する方法にはいくつかの方法がある．質問・疑問を入力する段階で分類を指定してもらう，質問・疑問を字句解析し分野を推測する，質問・疑問の分野を問う仕事を導入する，などの方法が考えられる．

この分類は蓄積された知識データを検索するときの手助けにもなる．

4.2 仕事数

仕事は，新たな質問・疑問が投稿された時や，回答が返されたとき，その回答が複数存在する時などに生成される．この仕事数はシステムで制御する必要がある．質問・疑問に対して十分妥当である回答が得られた時は，それを知識データとして蓄積し，新たな仕事の生成を止めなければならない．

この制御には妥当性評価の仕事の結果を用いる．妥当性があると評価された数がある一定数を越えた場合，知識データとして十分であると判断し，仕事の生成を止める．

4.3 操作の制限

例えば参加者に回答の妥当性評価の仕事が来た時，その回答が間違っていることを知っていた場合，その参加者は正否の判断だけではなく新たな回答を書きたいと思うかもしれない．または，回答の比較の仕事が来た時に，それらの回答はどちらもそれなりに良い回答であるが，それらの良い所を合わせて一つにするとより良い回答になると判断するかもしれない．

本システムではそのような複雑な判断を要求するような仕事を参加者に要求しない．しかし，もしそのような操作をしたいと考える参加者がいた場合，その操作を可能にするシステムは必要である．

そこで本システムには，与えられた仕事こなすことでシステムに参加するインタフェースの他に，積極的に仕事が行なえるようなインタフェースを用意する．例えば上記例であれば，それぞれ仕事から実際に回答をするための仕事を引き出せるインタフェースを用意すれば良い．

5 関連研究

人の力を借りる広域分散システムには The Worldwide Lexicon [5] や Open Mind Initiative [2], Wikipedia [6] などがある．

The Worldwide Lexicon は言語に関する知識を集め翻訳に使える辞書を作り上げるシステムである．このシステムでは SETI@home のようにクライアントプログラムを配布して，それを使って新たな語彙情報の追加をしてもらう．その追加情報は各地に広がる辞書サーバに蓄えられ，他のユーザが利用することが可能となる．

Open Mind Initiative はプロジェクトが運営する Web サーバ上でユーザにデータの評価をしてもらい，音声認識や認知などの情報を構築していくシステムである．Open Mind Initiative は一つの Web サーバのみでデータを公開しているため，多くのユーザ数を獲得することは困難である．

Wikipedia は誰でも自由に書き込み，書き換えが可能な Web ページを用いた百科事典構築プロジェクトである．本研究と同様，このシステムの利用を通じてインターネット上の知識データを増やすことが可能である．しかし，Wikipedia も Open Mind Initiative と同様参加者は単一の Web サーバ上での作業しか可能でないため，多くの参加者を獲得することはできていない．またこのシステムは質問・疑問を積極的に管理する機能は有していないため，知識データ生成の機会を．

6 おわりに

本稿では，現在の知識のデータ化プロセスを考察しその問題点を挙げ，その問題点を踏まえたシステムを提案した．

システム構造として SETI@home のような広域分散システムを採用し，質問・疑問を多くの人間に提示し，その情報を管理する機構を導入した．参加者に仕事として質問・疑問を配布し，その回答を回収することで，知識データの増加を促進する．本システムはシステム内部のプロトコルには SOAP を採用し，様々なシステムをインタフェースとして利用できる．またその他のサービスとの連係も可能である．

また知識データの品質を高める機構を導入した．知識データの妥当性を保障するため，回答の妥当性を問う仕事を参加者に配布する．また複数個の回答が返って来た場合，その優劣を比較する仕事を参加者に配布する．

その後，本システムにおける課題を挙げ，その解決法を示した．

今後はさらに考察を深めたのち，システムの実装，実際のシステム運用を通しての評価，などの展開が考えられる．

参考文献

[1] C. Kwok, O. Etzioni and D. Weld, Scaling Question Answering to the Web, In Proceedings of WWW10, Hong Kong, (2001).

[2] Open Mind Initiative, `http://www.openmind.org/index.shtml`.

[3] SETI@home, `http://setiathome.ssl.berkeley.edu/`.

[4] T. Box, D. Ehnebuske, G. Kakivaya, A. Layman, N. Mendelsohn, H. Nielson, S. Thatte, D. Winer, SOAP: Simple Object Access Protocol, `http://www.w3.org/TR/SOAP`.

[5] The Worldwide Lexicon, `http://picto.weblogger.com/`.

[6] Wikipedia, `http://www.wikipedia.com/`.