(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3896014号
(P3896014)
(45)発行日　平成19年3月22日(2007.3.22)　(24)登録日　平成18年12月22日(2006.12.22)

(51)Int.Cl.		F I		
G06F 17/30	(2006.01)	G06F	17/30	340A
G06F 13/00	(2006.01)	G06F	17/30	180Z
		G06F	17/30	380D
		G06F	13/00	560A

請求項の数 6　(全 24 頁)

(21)出願番号　特願2002-81642(P2002-81642)
(22)出願日　平成14年3月22日(2002.3.22)
(65)公開番号　特開2003-281173(P2003-281173A)
(43)公開日　平成15年10月3日(2003.10.3)
審査請求日　平成15年4月17日(2003.4.17)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　100058479
弁理士　鈴江　武彦
(74)代理人　100084618
弁理士　村松　貞男
(74)代理人　100092196
弁理士　橋本　良郎
(74)代理人　100091351
弁理士　河野　哲
(74)代理人　100088683
弁理士　中村　誠
(74)代理人　100070437
弁理士　河井　将次

最終頁に続く

(54)【発明の名称】情報収集システム、情報収集方法及びコンピュータに情報収集を実行させるプログラム

(57)【特許請求の範囲】

【請求項１】

ユーザの要求を満足する情報を収集して提示する情報収集システムにおいて、

それぞれ複数のユーザをメンバーとする複数のコミュニティを管理するコミュニティ管理手段と、

各コミュニティに属するメンバーがメッセージの送受信を行うためのメッセージ送受信手段と、

前記複数のコミュニティの各々で共有されている情報をユーザが閲覧するためのコミュニティ情報提示手段と、

各コミュニティに属するメンバーが共同で、当該コミュニティにおける収集要求として、収集の起点とする情報と、情報が含むべき語句の条件とを記述した収集要求を編集するための収集要求編集手段と、

各複数のコミュニティにおいて編集された複数の収集要求のいずれかを満足する情報を、前記各々の収集要求に記述された収集の起点である情報からハイパーリンクを辿って前記語句の条件を満たす情報を探索することにより、情報ネットワーク上の複数の情報源から収集する情報収集手段と、

前記収集した情報に基づいて前記複数の収集要求の各々に対応する収集結果を各々生成する収集結果生成手段と、

各コミュニティに属するメンバーが共同で、当該コミュニティにおける収集結果を編集するための収集結果編集手段と、を具備し、

前記コミュニティ情報提示手段は、複数のコミュニティで各々作成された複数の収集結果を、当該コミュニティまたは当該コミュニティ内で送受信されるメッセージと関連付けて、当該コミュニティのメンバーおよび非メンバーのユーザに提示すると共に、当該コミュニティにおける当該収集結果を構成する前記情報収集手段で収集された情報と、他のコミュニティにおける収集結果を構成する前記情報収集手段で収集された情報とが、重複する場合には、当該コミュニティにおける当該収集結果と、当該他のコミュニティにおける収集結果とを、関連付けて提示することを特徴とする情報収集システム。

【請求項２】

請求項１に記載の情報収集システムにおいて、コミュニティのメンバーが前記メッセージ送受信手段を用いて送受信するメッセージから、前記情報収集の起点とし得る情報と、当該情報に関わるコメント文とを抽出して、これらに基づき、当該コミュニティの収集要求及び当該コミュニティの収集結果の少なくとも一方を自動的に更新することを特徴とする情報収集システム。

【請求項３】

請求項１又は請求項２に記載の情報収集システムにおいて、コミュニティのメンバーが前記収集結果編集手段を用いて行った収集結果の編集内容に基づき、当該収集結果に対応する収集要求を更新することを特徴とする情報収集システム。

【請求項４】

請求項１から請求項３のいずれか１項に記載の情報収集システムにおいて、ユーザが入力する検索条件を満足する情報を、前記情報収集手段で収集した情報の中から検索する収集情報検索手段をさらに具備し、当該収集情報検索手段は、検索された情報と、コミュニティで作成した収集結果のうち前記検索された情報を含む収集結果とを、関連付けて提示することを特徴とする情報収集システム。

【請求項５】

ユーザの要求を満足する情報を収集して提示する情報収集方法において、

コンピュータが、各コミュニティに属するメンバーが共同で、当該コミュニティにおける収集要求として、収集の起点とする情報と、情報が含むべき語句の条件とを記述した収集要求を編集し、

コンピュータが、各コミュニティにおいて編集された複数の収集要求のいずれかを満足する情報を、前記各々の収集要求に記述された収集の起点である情報からハイパーリンクを辿って前記語句の条件を満たす情報を探索することにより、情報ネットワーク上の複数の情報源から収集し、

コンピュータが、前記収集した情報に基づいて前記複数の収集要求の各々に対応する収集結果を各々生成し、

コンピュータが、各コミュニティに属するメンバーが共同で、当該コミュニティにおける収集結果を編集し、

コンピュータが、複数のコミュニティが各々作成された複数の収集結果を、当該コミュニティまたは当該コミュニティ内で送受信されるメッセージと関連付けて、前記複数のコミュニティの各々で共有されている情報を、当該コミュニティのメンバーおよび非メンバーのユーザに提示すると共に、当該コミュニティにおける当該収集結果を構成する前記情報収集手段で収集された情報と、他のコミュニティにおける収集結果を構成する前記情報収集手段で収集された情報とが、重複する場合には、当該コミュニティにおける当該収集結果と、当該他のコミュニティにおける収集結果とを、関連付けて提示することを特徴とする情報収集方法。

【請求項６】

コンピュータにユーザの要求を満足する情報を収集して提示する情報収集を実行させるプログラムにおいて、

コンピュータに、各コミュニティに属するメンバーによって共同で編集された当該コミュニティにおける収集要求として、収集の起点とする情報と、情報が含むべき語句の条件とを記述した収集要求を入力させ、

コンピュータに、各コミュニティにおいて編集された複数の収集要求のいずれかを満足する情報を、前記各々の収集要求に記述された収集の起点である情報からハイパーリンクを辿って前記語句の条件を満たす情報を探索することにより、情報ネットワーク上の複数の情報源から収集させ、

コンピュータに、前記収集した情報に基づいて前記複数の収集要求の各々に対応する収集結果を各々生成させ、

コンピュータに、各コミュニティに属するメンバーが共同で編集された当該コミュニティにおける収集結果を入力させ、

コンピュータに、複数のコミュニティで各々作成された複数の収集結果を、当該コミュニティまたは当該コミュニティ内で送受信されるメッセージと関連付けて、前記複数のコミュニティの各々で共有されている情報を、当該コミュニティのメンバーおよび非メンバーのユーザに提示させると共に、当該コミュニティにおける当該収集結果を構成する前記情報収集手段で収集された情報と、他のコミュニティにおける収集結果を構成する前記情報収集手段で収集された情報とが、重複する場合には、当該コミュニティにおける当該収集結果と、当該他のコミュニティにおける収集結果とを、関連付けて提示させることを特徴とするプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、インターネットやイントラネットなどの情報ネットワークに分散して配置された複数の情報源から、ユーザの要求を満足する情報を収集する情報収集システムに関する。

【０００２】

【従来の技術】

大規模情報ネットワーク技術の普及により、誰もが大量の情報を自由に利用できるようになっている。インターネットやイントラネットなどの情報ネットワーク上には、大量の情報がハイパーテキスト形式の文書（ウェブページ）として公開されており、その数は数十億とも言われている。これらの情報を利用する方法として、ブラウザと呼ばれる情報閲覧ソフトウェアを用い、関心のある項目（ハイパーリンク）を選択していく（ブラウジングする）方法が用いられるのが一般的である。また、大量の情報の中から、キーワード等で指定した条件を満足する情報を検索するための検索サービスサイトや、情報を利用しやすい形に分類して提供するディレクトリサイトが、各種運用されている。ユーザは、所望の情報を得るためには、まず、検索サービスサイトやディレクトリサイトを用いて自分の関心に合致しそうな文書を求めた後に、その文書の内容や、その文書にリンクされた他の文書の内容を、ブラウジングすることによって調べるという、一連の作業を繰返し行う。また、頻繁に利用する情報や特に重要な情報については、ブラウザの付属機能であるブックマークと呼ばれる手段を用いてその情報の位置（ＵＲＬ）を記憶したり、有用な情報の位置をリストアップした文書（リンク集）を作成して利用することが行われている。

【０００３】

しかし、大量の情報の中から、検索やブラウジングによって必要な情報を集める作業は時間と労力を要する。また、検索サービスサイトやディレクトリサイトでは、最新の情報や、専門性の高い情報が十分に提供されていないという問題もある。これらの問題を解決する技術の一つに、自動クローリング技術が知られている。これは、ハイパーテキストのハイパーリンクを再起的に辿る（すなわち、クローリングする）ソフトウェア（すなわち、クローラ）を用いて、大量の文書情報を自動的に走査し、ユーザが指定した条件を満足する文書を収集する方法である。ユーザがクローラに与えることのできる収集条件には、収集する文書の個数・容量の制限や、収集を開始する起点の文書、起点の文書から辿るリンクの段数の上限、収集する範囲（ウェブサーバのドメインなど）、文書の更新日時の条件、などがある。また、文書の内容に関する条件としては、キーワード・フレーズ等が対象文書中で出現する頻度や、例示した文書と対象文書との類似度、ユーザの興味・関心の記

述（プロファイル）と対象文書との類似度、などについての条件がある。さらには、対象文書の重要度を、アクセス数やハイパーリンクの構造に基づいて計算し、重要度の大きい文書を優先的に収集する方法なども提案されている。自動クローリング技術に関する公知文献には、"Focused Crawling: A New Approach for Topic-Specific Resource Discovery", Soumen Chakrabarti他, The Eighth International World Wide Web Conference, 1999（以下、「文献１」と称する）や、特開平１０－２６０９７８号公報「情報収集方法及び装置」（以下、「文献２」と称する）などがある。

【０００４】

一方、複数のユーザが互いに情報を交換するための手段としては、電子メールおよびメーリングリスト、電子掲示板、チャットなどの手段が、広く普及している。メーリングリストは、複数のユーザの電子メールアドレスをまとめて、その全員に一括してメッセージを送信できるようにした手段である。また、電子掲示板は、ネットワーク上に情報共有のためのスペースを設けて、複数の登録ユーザあるいは匿名ユーザが自由にメッセージを記入できるようにした手段である。チャットは、電子掲示板と同様に情報共有スペースを設けて、テキストのメッセージをリアルタイムに送受信できるようにした手段である。メーリングリストや電子掲示板、チャット等のように、比較的多数のユーザによる（一対一のみでない）メッセージの交換を目的としたコミュニケーション手段では、参加メンバーの大部分が共通に関心を持つ話題に関するメッセージがやり取りされることが多い。このように、共通の目的や話題を持って電子的なメッセージを交換するユーザの集団を、本明細書においては、以下、「コミュニティ」と称する。

【０００５】

コミュニティのメンバーの一人が有用な情報を得た場合、上述のコミュニケーション手段を用いて他のメンバーに通知することによって、メンバー間で情報を共有するということが日常的に行われている。このようにして交換される情報のうち、とくに有用な情報については、メンバーの有志が自発的に、有用な情報を手作業でリストアップし、他のメンバーが利用しやすいようにリンク集などの形に整理し、定期的に保守するということが行われる場合もある。コミュニティのメンバーが関心を持つ話題は、コミュニティの趣旨を逸脱しない範囲内にある場合が多いが、多少は動的に変遷する。コミュニティのメンバーがどのような話題に関心を持っているかを自動的に調べる技術については、特開２０００－２９３５２６号公報「嗜好情報収集システム」（以下、「文献３」と称する）や、特開２００１－９２７５５号公報「プロファイル作成方法及びシステム」（以下、「文献４」と称する）などの公知文献がある。

【０００６】

【発明が解決しようとする課題】

自動クローリングは、収集に要する時間とネットワーク資源の消費が大きいわりに収集の効率が良くないという問題がある。インターネットからのクローリングによる収穫率、すなわち、収集したウェブページの中にユーザの要求と関連する情報が含まれる割合は、最良の場合で50%程度とされており（文献１）、残りの50%のページは利用されずに捨てられることになる。文献１と文献２では、収集の効率を改善するための方法が開示されているが、そもそもインターネット上には、有用でない情報も多数含まれている。例えば、ユーザの収集要求をキーワード集合で記述した場合、そのキーワード集合を多く含んだ文書でさえ、ユーザにとって実際に有用であるとは限らず、古い情報や誤った情報、冗長な情報である可能性がある。したがって、収集効率の改善には限界があり、収集された情報が有用かどうかの判断はユーザに委ねざるを得ない。また、個々のユーザが個別にクローラを利用することは、通信ネットワークやプロキシサーバ、ウェブサーバなどにかかる負荷が大きくなるため、現実的でない。従って、より効率的な収集方法と、収集結果を無駄にせずに再利用する方法が望まれる。

【０００７】

さらに、クローリングによってウェブページを収集するには、収集の条件として、収集を開始する起点のＵＲＬや収集する範囲、キーワードなどの条件をユーザが指定する必要が

ある。しかし、どのような条件を指定すれば有用な情報が得られるかが不可知である上、上述のように収集効率が良くない。従って、一般的に、検索サービスサイトや、配信型の情報フィルタリングシステムと比較して、クローラを利用するには熟練を要する。このため、有用な情報を効率よく収集するための知識やノウハウをユーザ間で共有することが望まれる。

【０００８】

以上のような理由のため、クローラは、主に、検索サービスサイトが、任意の内容のウェブページを大量に収集してインデキシングする目的と、既知の限定されたウェブサイトを定期的に巡回して、更新された情報の有無を監視する目的に利用されるにとどまっている。従って、クローラが、未知の情報源から積極的に情報を収集したり、潜在的にユーザの関心に合致するであろう新しい情報を発見したりする目的に活用されていないのが現状である。

【０００９】

一方、コミュニティのメンバーが電子掲示板等の従来のコミュニケーション手段を用いて情報をやり取りする方法では、メンバー各々の知識や専門性を生かした情報の共有を柔軟に行うことができる。しかしこの方法は、個々のユーザの能力と自発性に依存するところが大きい。有用な情報を探して他のメンバーに知らせる作業は労力を要するし、そもそも、コミュニティのメンバー全員が知らないような新しい情報を発見することは不可能である。文献３と文献４には、コミュニティでやり取りされるメッセージを解析して、ユーザの関心や嗜好（プロファイル）を求める発明が開示されているが、これらの発明は、コミュニティのメンバーの関心・嗜好に合った情報を新たに収集する手段を提供するものではない。

【００１０】

また、有用な情報が個々のメンバーの努力によって数多く得られたとしても、その各々が未整理のまま別々のメッセージに分散している状態では、収集した情報を有効活用することができない。有用な情報を大量のメッセージの中から選び出してコミュニティのメンバー間で共有できる形に整理する作業には労力を要するが、その作業もメンバー各々の自発的な手作業に負っている。文献３に係る発明はユーザの嗜好調査、文献４に係る発明は、ユーザを関心・嗜好に基づいてカテゴライズした結果を明示することにより、コミュニケーションの円滑化を図ることを目的とする。いずれの発明も、コミュニティのメンバーのために有用な情報を整理したり保守するという作業を支援するものではない。

【００１１】

本発明は、上記の課題を解決するためになされたものであり、ユーザの要求を満足する情報を効率よく収集するとともに、その収集結果を複数のユーザで有効に活用し、かつ、有用な情報を継続的に整理・保守する作業を支援することを目的とする。

【００１２】

【課題を解決するための手段】

前記課題を解決するために、本発明に係る情報収集システムは、ユーザの要求を満足する情報を収集して提示する情報収集システムにおいて、それぞれ複数のユーザをメンバーとする複数のコミュニティを管理するコミュニティ管理手段と、各コミュニティに属するメンバーがメッセージの送受信を行うためのメッセージ送受信手段と、前記複数のコミュニティの各々で共有されている情報をユーザが閲覧するためのコミュニティ情報提示手段と、各コミュニティに属するメンバーが共同で、当該コミュニティにおける収集要求を編集するための収集要求編集手段と、各複数のコミュニティにおいて編集された複数の収集要求のいずれかを満足する情報を情報ネットワーク上の複数の情報源から収集する情報収集手段と、前記収集した情報に基づいて前記複数の収集要求の各々に対応する収集結果を各々生成する収集結果生成手段と、各コミュニティに属するメンバーが共同で、当該コミュニティにおける収集結果を編集するための収集結果編集手段と、を具備し、前記コミュニティ情報提示手段は、複数のコミュニティで各々作成された複数の収集結果を、当該コミュニティまたは当該コミュニティ内で送受信されるメッセージと関連付けて、当該コミュ

ニティのメンバーおよび非メンバーのユーザに提示することを特徴とする。
【００１３】
本発明に係る情報収集システムの好ましい実施態様は以下のとおりである。なお、以下の各実施態様は、単独で適用しても良いし、適宜組み合わせて適用しても良い。
【００１４】
（１） コミュニティのメンバーが前記メッセージ送受信手段を用いて送受信するメッセージに基づき、当該コミュニティの収集要求及び当該コミュニティの収集結果の少なくとも一方を自動的に更新すること。
【００１５】
（２） コミュニティのメンバーが前記収集結果編集手段を用いて行った収集結果の編集内容に基づき、当該収集結果に対応する収集要求を更新すること。
【００１６】
（３） コミュニティの収集結果と、当該コミュニティの収集結果に含まれる情報を重複して含む他のコミュニティの収集結果とを関連付けて提示すること。
【００１７】
（４） ユーザが入力する検索条件を満足する情報を、前記情報収集手段で収集した情報の中から検索する収集情報検索手段をさらに具備し、当該収集情報検索手段は、検索された情報と、コミュニティで作成した収集結果のうち前記検索された情報を含む収集結果とを、関連付けて提示すること。
【００１８】
本発明に係る情報収集方法は、ユーザの要求を満足する情報を収集して提示する情報収集方法において、各コミュニティに属するメンバーが共同で、当該コミュニティにおける収集要求を編集し、各コミュニティにおいて編集された複数の収集要求のいずれかを満足する情報を情報ネットワーク上の複数の情報源から収集し、前記収集した情報に基づいて前記複数の収集要求の各々に対応する収集結果を各々生成し、各コミュニティに属するメンバーが共同で、当該コミュニティにおける収集結果を編集し、複数のコミュニティが各々作成された複数の収集結果を、当該コミュニティまたは当該コミュニティ内で送受信されるメッセージと関連付けて、前記複数のコミュニティの各々で共有されている情報を、当該コミュニティのメンバーおよび非メンバーのユーザに提示することを特徴とする。
【００１９】
本発明に係る情報収集コンピュータにユーザの要求を満足する情報を収集して提示する情報収集を実行させるプログラムは、コンピュータにユーザの要求を満足する情報を収集して提示する情報収集を実行させるプログラムにおいて、各コミュニティに属するメンバーによって共同で編集された当該コミュニティにおける収集要求を入力し、各コミュニティにおいて編集された複数の収集要求のいずれかを満足する情報を情報ネットワーク上の複数の情報源から収集し、前記収集した情報に基づいて前記複数の収集要求の各々に対応する収集結果を各々生成し、各コミュニティに属するメンバーが共同で編集された当該コミュニティにおける収集結果を入力し、複数のコミュニティで各々作成された複数の収集結果を、当該コミュニティまたは当該コミュニティ内で送受信されるメッセージと関連付けて、前記複数のコミュニティの各々で共有されている情報を、当該コミュニティのメンバーおよび非メンバーのユーザに提示することを特徴とする。
【００２０】
【発明の実施の形態】
以下、図面を用いて本発明の実施の形態を説明する。
【００２１】
図１は、本発明の一実施形態に係る情報収集システムの構成を示す図である。図１において、コミュニティ管理部１は、複数のコミュニティを管理する。すなわち、コミュニティ管理部１は、複数のコミュニティの各コミュニティにおけるメンバーである複数のユーザの情報と、各コミュニティにおけるユーザ間で送受信されるメッセージとを記憶管理する。コミュニティ管理部１は、従来技術による電子掲示板あるいはメーリングリスト等の管

理手段と同様に、ユーザ情報記憶部１１とメッセージ記憶部１２を有する。通常、コミュニティのメンバーと非メンバーでは、アクセス権、すなわち、ユーザ情報の閲覧やメッセージの送受信などが行えるか否かの権限が異なるが、このコミュニティ管理部１が、そのアクセス制御を行う。また、本明細書においては、ユーザは、メンバーと非メンバーを含むものとする。また、詳細は後述するように、コミュニティ管理部１は、ユーザからの情報収集の要求を複数記憶する収集要求記憶部１３と、情報収集の結果としてユーザに提示する情報を複数記憶する収集結果記憶部１４とを有する。

【００２２】

コミュニティ情報提示部２は、複数のコミュニティの名称やメンバーなどの基本的な情報や、個々のコミュニティ内でやり取りされるメッセージや共有文書などの情報をユーザに提示する。これにより、ユーザが様々な情報を閲覧できる。

【００２３】

メッセージ送受信部３は、コミュニティのメンバーが、他のメンバーに対してメッセージを送信・受信するための手段である。メッセージ送受信部３で送受信されたメッセージは、コミュニティ毎に整理されて、メッセージ記憶部１２に記憶される。

【００２４】

収集要求編集部４は、情報収集の要求をコミュニティの複数のメンバーが共同で編集して登録するための手段であり、収集要求編集部４で編集された結果は、収集要求記憶部１３にコミュニティ毎に記憶される。同様に、収集結果編集部５は、情報収集の結果をコミュニティの複数のメンバーが利用しやすい形に編集するための手段であり、収集結果編集部５で編集された結果は、収集結果記憶部１４に、コミュニティ毎に記憶される。

【００２５】

情報収集部６は、収集要求記憶部１３に記憶された複数の収集要求を入力として、インターネットやイントラネットなどの情報ネットワークから、いずれかの収集要求を満足する情報（本実施形態の場合はウェブ文書）を収集する。情報収集部６で収集されたウェブ文書は、ウェブ文書記憶部７にインデキシングされて記憶される。

【００２６】

収集結果生成部８は、コミュニティ毎に登録された収集要求に基づき、収集したウェブ文書から要求に合致するものを選択・加工して、コミュニティ毎に収集結果を生成する。この収集結果は収集結果記憶部１４に記憶されるが、ユーザは、必要に応じて収集結果編集部５を用いて収集結果をより利用しやすい形に編集して保存することができる。

【００２７】

以上に説明した構成は、本発明を実施するための最小の構成であるが、上記の構成に加え、さらに、収集要求生成部９を備えてもよい。収集要求生成部９は、個々のコミュニティのメンバーが送受信するメッセージに基づき、当該コミュニティの収集要求を自動的に生成あるいは追加する。これと同様に、収集結果生成部８に、メッセージに基づいて収集結果を生成あるいは追加する機能を持たせることも可能である。さらに、収集結果生成部８に、ユーザが収集結果を編集した場合に、その編集内容に基づいて、対応する収集要求を変更する機能を持たせることも可能である。

【００２８】

ウェブ文書検索部１０は、情報ネットワークから収集して前記のウェブ文書記憶部７に記憶したウェブ文書を、ユーザが検索して利用するための手段である。ウェブ文書検索部１０の手段の検索機能は、従来技術によるウェブ文書の検索手段と概ね同じである。本発明の実施形態に係るウェブ文書検索部１０は、検索結果を提示する際に、収集結果記憶部１４に記憶されている収集結果を併せて提示する処理を行う機能を備えている。

【００２９】

以上に説明した本発明の実施形態に係る情報収集システムの構成と、従来の情報収集システムの構成との違いを、図２を参照して説明する。図２は、従来の一般的な情報収集システムの概略ブロック図である。図２に示す情報収集システムは、図１の構成要素でもある、収集要求編集部４、収集要求記憶部１３、情報収集部６、ウェブ文書記憶部７、収集結

果生成部８、収集結果記憶部１４、収集結果編集部５、および、場合によりウェブ文書検索部１０を具備している。しかしながら、従来の情報収集システムは、収集要求の作成から収集結果の作成、編集までを一人のユーザが行うように構成されている。このため、従来の情報収集システムは、複数のユーザ、すなわちコミュニティで協力しあって情報を収集する目的には利用できない。また、従来の情報収集システムは、収集された情報や新たに収集すべき情報についての議論や情報交換といった活動を行うための手段も具備せず、加えて、収集結果を複数のユーザで共有し保守するための手段も備えていない。このような構成では、ユーザの労力が大きいだけでなく、複数のユーザによる情報収集結果の共有・再利用が行えないという問題がある。

【００３０】

以下、本発明の実施形態について詳細に説明する。

【００３１】

図３は、ユーザ情報記憶手段に記憶されるユーザの情報とコミュニティの情報を示す図である。図３（ａ）は、ユーザ情報３１の一例であり、図３（ｂ）は、コミュニティ情報３２の一例である。ユーザ情報３１は、本情報収集システムを利用する個々の登録ユーザ（所定の権限が与えられている既知のユーザ）の情報であり、ユーザＩＤ、パスワード、氏名、メールアドレス、所属コミュニティ、ホームページＵＲＬ等の項目を有する。コミュニティ情報３２は、本情報収集システムが管理するコミュニティの情報であり、コミュニティＩＤ、コミュニティ名、メーリングリストアドレス、掲示板ＵＲＬ、および、参加メンバーのユーザＩＤ等の項目を有する。メーリングリストアドレスは、コミュニティのメンバー全員に一括してメッセージを送信する際の宛先である。掲示板ＵＲＬは、メッセージを記入して共有するスペースの位置を表す。メーリングリストアドレスと掲示板ＵＲＬのどちらか一方が設定されていれば、その手段を用いてメンバー間のメッセージの交換が可能となる。メーリングリストアドレスと掲示板ＵＲＬの両方が設定されていれば、ユーザは、利用しやすい方の手段を利用することができる。また、コミュニティ情報３２のコミュニティのメンバーの項目は、ユーザ情報３１のユーザＩＤによって記述される。逆に、ユーザ情報３１の所属コミュニティの項目は、コミュニティＩＤによって記述される。

【００３２】

ユーザがコミュニティを利用して情報交換の作業を行うための手順を、図４のフローチャートを用いて説明する。まず、ユーザが登録ユーザなら（ステップ４１）、ユーザ認証を行う（ステップ４２）。ステップ４２で、認証に成功すれば（ステップ４３）、当該登録ユーザの権限でコミュニティを利用できるようになる。ユーザ認証の手続きは、従来の方法と同じく、ユーザが入力したユーザのＩＤとパスワードを認証する方法でよい。ユーザが未登録のユーザであり、かつ新規にユーザ登録を希望するなら（ステップ４４）、ユーザ登録手続き（ステップ４５）を行う。ステップ４５で、登録が正しく行えたならば（ステップ４６）、新規の登録ユーザとしての権限でコミュニティを利用できるようになる。ユーザの登録の手続きは、従来の方法と同様に、図３（ａ）に示すユーザ情報３１を、ユーザＩＤ３１のうち氏名、パスワード等の必須項目をユーザに入力させ、ユーザＩＤを新しく発行することによってなされる。以上の処理は、コミュニティ管理部１が行う。

【００３３】

その後、コミュニティ情報提示部にて、既存のコミュニティの一覧をユーザに提示する。まず、登録ユーザに対してのみ、当該ユーザが参加しているコミュニティの一覧を提示する（ステップ４７）。次に、登録ユーザと、未登録の匿名ユーザの両方に対して、非参加のコミュニティの一覧を提示する（ステップ４８）。ここで、コミュニティに参加していないユーザや匿名ユーザについては、制限した情報のみを提示する。図５は、複数のコミュニティ情報の一覧を提示例した画面の例を示す。この画面は、登録ユーザ「ａｏｋｉ」（図３のユーザＩＤがｕ１のユーザ）がユーザ認証を行った段階において提示される「ａｏｋｉのポータルページ」５１を示す図である。図５において、ユーザが参加しているコミュニティの一覧（図４のステップ４７）を示す部分５２には、「ｅコマース調査グループ」５３や、「プロ野球ファンの集い」などの参加コミュニティが提示される。各々のコ

ミュニティに関連付けて「新着メッセージ」５４、「新着情報」５４の情報が提示される。新着メッセージとは、当該コミュニティに新しく送信されたメッセージの一覧である。また、新着情報とは、後述する情報収集の処理によって新しく収集された情報である。このように、コミュニティ情報の一覧画面では、コミュニティ毎に、メンバーが注目すべき新しい情報が明示的に提示される。一方、非参加コミュニティ５６とは、ユーザ「ａｏｋｉ」が参加していないコミュニティであり、「Ｌｉｎｕｘユーザ会」５７、「ガーデニング倶楽部」などのコミュニティが提示される。非参加コミュニティに関連付けて表示される「トピック」５８は、当該コミュニティが関心を持って情報を収集しているトピック（話題）を表すものであり、例えば「Ｌｉｎｕｘユーザ会」５７は「Ｌｉｎｕｘ」「ディストリビューション」などのトピックに関心を持つコミュニティであることが、非参加ユーザにも提示される一方で、具体的なメッセージなどの情報は非参加ユーザには提示されないようになっている。以上に説明したコミュニティ情報の提示の処理は図１のコミュニティ情報提示部２にて行われる。

【００３４】

次に、ユーザが、ある一つのコミュニティを選択してこれに加わり、メッセージ送受信などの活動をおこなうための処理の流れを説明する。ユーザが、図４のステップ４９にて選択したコミュニティに入ると、まず、ユーザがコミュニティのメンバーかどうかが確認される（ステップ４１０）。ステップ４１０において、ユーザが、当該コミュニティのメンバーでなく、かつ、コミュニティへの新規参加をユーザが希望するなら（ステップ４１１）、コミュニティへの加入手続きを行う（ステップ４１２）。この際、前記の登録ユーザのみがコミュニティへの加入対象となり、匿名ユーザはコミュニティへの参加は行えない。この加入手続き（ステップ４１２）は、図３（ｂ）に示したコミュニティ情報３２のメンバーの項目に、新規ユーザのＩＤを加えることによってなされるが、コミュニティの管理者や他のメンバーによって加入の可否を決定する手続きを含めてもよい。コミュニティのメンバーは、当該コミュニティ内でのメッセージの送受信と、収集要求・収集結果の閲覧・編集などの活動を行うことができる（ステップ４１４）。一方、コミュニティのメンバーでないユーザや、匿名ユーザは、コミュニティの利用が制限付きで許される（ステップ４１５）。図４に示した例では、非メンバーは、メッセージと収集結果の閲覧のみ許可され、編集は禁止するとして説明したが、コミュニティの性質によっては、これと異なるかたちで権限を許可あるいは禁止してもよい。ユーザは、コミュニティ内での活動を適宜行った後、コミュニティから出て（ステップ４１６）終了したり（ステップ４１７）、他のコミュニティに入って活動を行ってもよい。なお、図４では省略するが、コミュニティからの脱退やユーザ情報の変更、コミュニティの新規作成など、従来のコミュニティ管理システムで備えるべき処理機能も本発明の実施形態に係る情報収集システムは具備するものとする。さらに、本実施形態では主に、従来技術での電子掲示板と類似した画面例を用いて説明するが、メーリングリストのような手段を用い、ユーザ登録やコミュニティへの参加、情報の閲覧などの処理を電子メールで行うことも可能である。

【００３５】

図６から図８は、コミュニティ内でのユーザの活動として、メッセージの送信、収集要求の編集、収集結果の編集の処理の流れを示した図である。また、図９から図１３は、これらの処理に対応する画面の表示例である。メッセージの送受信は、本実施形態の場合は電子掲示板用の手段を用い、図１のメッセージ送受信部３にて行う。図９に示した画面例では、コミュニティ「ｅコマース調査グループ」に入ったユーザが、掲示板のメニュー９１を選択すると、画面上には最近に送信されたメッセージ９２、９４、９５等が表示される。個々のメッセージの間には返信関係が付されており、たとえばメッセージ９５とメッセージ９６は、ともにメッセージ９４の返信メッセージである。画面上でユーザが一つのメッセージを選択すると、その内容が提示されるようになっており、例えば図９では、ユーザが選択したメッセージ９６「著名なオークションサイト」（１月１０日にユーザｙａｍａｄａによって送信されたメッセージ）の内容として、テキスト９７が表示された場面を示している。メッセージのうち、後述する情報収集の結果と関連するものについては、こ

れらが互いに関連付けられて提示される。例えば図９では、メッセージ９２「音楽配信ビジネス」に対して、このコミュニティが共同で情報収集を行っているトピック「コンテンツ配信」９３が関連付けられて表示される。

【００３６】

一方、メッセージの送信は、図６で示した処理により行われる。ユーザは、まずステップ６１で、送信したいメッセージが既存メッセージの返信であるか、あるいは、新規のメッセージであるかを選択する。この選択は、図９の画面例ではボタン９８またはボタン９９を押すことによって行う。ここで、「返信メッセージ」のボタン９８を押した場合は、図９で現在表示しているメッセージ９６に対する返信メッセージを作成することになる。図１０は返信メッセージの作成（図６のステップ６２）の画面例を表す図である。ユーザは、図１０に示す返信メッセージのタイトル１０１と本文１０２とを、必要ならば返信元のメッセージを引用して編集し、返信メッセージを作成する。その後、既存メッセージの返信メッセージとして送信する（ステップ６３）と、上記に説明した返信関係が付されてシステムに記憶される。メッセージの送信は、図１０に示す「送信」ボタン１０３を押すことによって行われる。新規メッセージの場合も、既存メッセージの返信メッセージの送信と同様に、図６のステップ６４、ステップ６５の処理を経て送信が行われる。送信されたメッセージは図１のメッセージ記憶部１２に記憶され、コミュニティのメンバーが図９で説明した形で閲覧したり、新しいメッセージを返信したりすることができるようになる。

【００３７】

図７は、収集要求をコミュニティのメンバーが編集する処理の流れを示す図である。収集要求は、本明細書においては、ユーザが、どのような情報を収集したいかの要求や条件を記述したデータをいい、図１の情報収集部６の入力となる。本実施形態では、収集要求をコミュニティの複数のメンバーが共同で編集することとしているため、編集内容の整合性を保つ必要がある。このため、まず、既に収集要求が存在するかどうかを確認する（ステップ７１）。ステップ７１において、収集要求が存在していない場合には、新規の収集要求を作成する（ステップ７６）。ステップ７１において、既に収集要求が存在する場合には、収集要求が他のユーザにチェックアウトされていないことを確認する（ステップ７２）。この確認後に、ユーザが収集要求を編集できるようになる。ステップ７２において、収集要求が他のユーザにチェックアウトされていなければ（ステップ７２のＹｅｓ）、まず、編集対象の収集要求が当該ユーザにチェックアウトされる（ステップ７３）。そして、ユーザによる編集作業（ステップ７４）の後に、チェックイン（ステップ７５）を経て、システムへの登録（ステップ７７）が行われる。なお、ステップ７２において、収集要求が他のユーザにチェックアウトされていれば（ステップ７２のＮｏ）、当該ユーザの収集要求は編集できないので、そのまま終了する。

【００３８】

以上説明した収集要求編集処理は、図１の収集要求編集部４にて行われ、編集された結果は収集要求記憶部１３に記憶される。なお、編集された収集要求は、過去の収集要求と置き換えて記憶してもよいし、過去のリビジョンを保存しておいて、編集毎に新たな収集要求を追加記憶してもよい。

【００３９】

図１１には、収集要求を編集する画面の例を示す。ユーザが画面上で収集要求のメニュー１１１を選択すると、収集要求を編集するための手段が表示される。コミュニティ内で収集を行いたいトピックは、通常複数あると考えられるので、一つのコミュニティが作成する収集要求の中で、複数のトピックを記述することができるようにしている。

【００４０】

図１１の例では「ｅコマース調査グループ」というコミュニティの収集要求の例として、「電子モール」「コンテンツ配信」「オンライン・トレード」のトピックが示されている。ユーザは、これらの既存のトピックの他に新しいトピックを追加したり（ボタン１１６）、不要となったトピックを削除したり（ボタン１１３）といった編集も可能である。なお、図７で説明したチェックアウト・チェックインの処理単位は、収集要求全体を１つ

の処理単位とするのでなく、トピックを１つの処理単位としてもよい。個々のトピック毎に記述するデータとしては、図１１に示すように、トピックの名称１１２、キーワード１１４、収集起点ＵＲＬ１１５がある。キーワード１１４は、収集した情報（本実施形態の場合はウェブ文書）がその内容に含むべきキーワードの論理式を記述する項目である。また、収集起点ＵＲＬは、クローリングを開始するウェブ文書のＵＲＬを記述する項目である。収集起点ＵＲＬは、必ずしも設定する必要はない。なぜならば、あるトピックの収集起点ＵＲＬが未指定であっても、複数のコミュニティが複数のトピックに記述した収集起点ＵＲＬのいずれかからクローリングすることによって、ユーザが所望する当該トピックの情報が収集できる可能性が高いからである。また、場合によっては、デフォルトの収集起点ＵＲＬとして、代表的なディレクトリサイト等を選ぶことにしてもよい。以上説明した項目を図１１の画面上で編集した後、「登録」ボタン１１７を押すことによって、編集後の収集要求がシステムに登録される。

【００４１】

図８は、収集結果をコミュニティのメンバーが編集する処理の流れを示す図である。収集結果は、情報要求に応じてシステムが収集した情報を、コミュニティのメンバーが利用しやすい形式に加工したデータをいい、主には図１の収集結果生成部８の出力である。収集結果は、必ずしもクローリングによって収集した情報のみからなるわけでなく、ユーザが明示的に有用と思う情報を記述してもよいし、後述するように、コミュニティのメンバー間で送受信されるメッセージに含まれる情報を追加してもよい。本実施形態では、前述の収集要求と同様に、収集結果もコミュニティの複数のメンバーが共同で編集することとしているため、編集内容の整合性を保つ必要がある。このため、まず、既に収集結果が存在するかどうかを確認する（ステップ８１）。ステップ８１において、収集結果が存在していない場合には、新規の収集結果を作成する（ステップ８６）。ステップ８１において、既に収集要求が存在する場合には、収集結果が他のユーザにチェックアウトされていないことを確認する（ステップ８２）。この確認後に、ユーザが編集できるようになる。ステップ８２において、収集結果が他のユーザにチェックアウトされていなければ（ステップ８２のＹｅｓ）、まず、編集対象の収集結果がチェックアウトされる（ステップ８３）。そして、ユーザによる編集作業（ステップ８４）の後に、チェックイン（ステップ８５）を経て、システムへの登録（ステップ８７）が行われる。なお、ステップ８２において、収集結果が他のユーザにチェックアウトされていれば（ステップ８２のＮｏ）、当該ユーザの収集結果は編集できないので、そのまま終了する。

【００４２】

以上説明した収集結果編集処理は、図１の収集結果編集部５にて行われ、編集された結果は収集結果記憶部１４に記憶される。図１２には、収集結果を表示する画面の例を示す。ユーザが画面上で収集結果のメニュー１２１を選択すると、収集結果を表示するための手段が表示される。収集結果は、上述の収集要求のトピック毎に整理されて表示される。図１２の例では、「ｅコマース調査グループ」の収集結果として、「電子モール」１２２、「コンテンツ配信」１２６等のトピック毎に整理されて情報が表示されている。さらに、個々のトピック中の情報は、サイト別に整理される。サイトは、インターネットにおける情報サービスの主体であり、情報源の単位でもある。図１２の例では、トピック「電子モール」１２２の中にサイト「○○モール」１２３が分類されている。テキスト１２４は、「○○モール」１２３を説明するコメント文であって、コミュニティのメンバーが当該サイトの内容を理解しやすいように、メンバーの一人または複数が共同で作成したテキストである。個々のサイトの中で特に有用な情報や、新しい情報については、図１２に示したように、サイト内の詳細情報１２５として提示する。

【００４３】

クローリングによる情報収集の結果としては、このような既知のサイト内の情報が収集される場合（図１２の情報１２５参照）と、新しいサイトが収集される場合（図１２の情報１２８の例）がある。後者の場合、新しいサイトを説明するテキストはまだユーザによって作成されていないため、当該サイトのウェブ文書のテキストがそのまま提示される（

図１２の情報１２９参照）が、これをより理解しやすいコメント文に直す必要がある。また一般に、クローラによって収集された情報は全てが有用な情報とは限らず、コミュニティのメンバーが共有するに値する情報を取捨・整理する作業が必要である。収集結果編集部５は、この作業をコミュニティの複数のメンバーが行うために設けられた手段であり、図１３は収集結果を編集するための画面の例である。

【００４４】

ユーザが図１２で示した画面上の「編集」ボタン（１２１０）を押すと、図１３に示すような画面が表示される。収集結果は上述のように、複数のトピック（「電子モール」１３１等）によって整理され、さらにトピックは、サイト（「○○モール」１３４等）によって整理される。ユーザは、新しいトピックの追加と不要なトピックの削除を行うことができる（図１３のボタン１３１１、１３３）。さらに、新しいサイトの追加と不要なサイトの削除を行うことができる（図１３のボタン１３２、１３６）。個々のサイト毎に編集すべき項目としては、サイト名１３４、サイトのＵＲＬ１３５、サイトを説明するためのコメント文１３７、および、サイト内の詳細情報１３８である。このうち、クローリングによる情報収集で自動的に獲得できないデータはコメント文なので、ユーザの編集作業としては、コメント文を作成することが主な作業の一つであるが、これは、当該サイトのウェブ文書から取得したテキストをもとに作成すればよい。その他の作業としては、サイトや詳細情報を取捨して不要なものを削除する作業が主となる。

【００４５】

以上の説明では、ユーザがコミュニティ内で行う活動と、そのために提供された本発明の実施形態に係る手段を中心に説明したが、以下は、ユーザが要求する情報を情報ネットワークから収集してユーザの要求に合った収集結果を生成する処理について説明する。図１４は、図１の情報収集部６が行う処理の流れを表す図である。また、図１４の処理の複数のステップから、収集した情報を収集結果に加える処理である図１５の処理が呼び出されるが、これは図１の収集結果生成部８が行う処理である。

【００４６】

情報収集部６は、収集対象の候補であるＵＲＬの集合を保持し、その個々のＵＲＬについて、ウェブ文書を既に取得したかどうかに係る情報や、最後に取得した日時、当該ＵＲＬのリンク元ＵＲＬおよびそのリンクのアンカーテキストの情報を、図１のウェブ文書記憶部７に記憶する。このＵＲＬ集合をＵとする。また、全コミュニティが作成する収集要求の集合をＲとする。

【００４７】

まず、Ｕの初期値を空集合とする（ステップ１４１）。その後、Ｒに新しい収集要求ｒが作成されるたびに、個々のｒのトピックの収集起点ＵＲＬとして新しいＵＲＬが登録されたかどうかをチェックする（ステップ１４２）。新しいＵＲＬｕ（以下、単に、「ｕ」とのみ表記する）が登録されれば、そのスコアを計算する（ステップ１４３）。ここで、ｕの、ある収集要求ｒに対するスコアｓ（ｕ，ｒ）は、次式で計算する。

【００４８】

【数１】

$$s(u,r) = \alpha * \sum_{v \to u} s(v,r) + \beta * \sum_{a:v \to u} sim(a,r) + \gamma * sim(du,r)$$

【００４９】

ここで、α、β、γは定数である。ｖはＵに含まれるＵＲＬ（以下、単に、「ｖ」とのみ表記する）であり、かつ、ｖはｕのリンク元であるとする。ｓ（ｖ，ｒ）はｖの収集要求ｒに対するスコアである。また、ａ：ｖ→ｕはｖからｕへのリンクに付されたアンカーテキストである。ｓｉｍ（ａ，ｒ）は、アンカーテキストａと収集要求ｒのキーワード集合との類似度である。ｄｕはｕのウェブ文書のテキストである。ｓｉｍ（ｄｕ，ｒ）はｄｕ

のテキストと収集要求ｒのキーワード集合との類似度である。収集要求ｒのキーワード集合とは、収集要求ｒの全てのトピックに記述されたキーワードの論理式に出現する（否定表現以外の）すべてのキーワードである。テキストｔとキーワード集合との類似度は、キーワードｋの重みｗｋにテキストｔ中のｋの頻度ｆ（ｔ，ｋ）を乗じた値を、キーワード集合の個々の要素について合計をとった値として計算する。すなわち、

【数２】

$$sim(t,r)=\sum_{k\in r} wk*f(t,k)/nr$$

とする。ｎｒは収集要求ｒのキーワード集合の要素数である。キーワードの重みｗｋはＩＤＦ（Inverted Document Frequency：すなわち、より多くのテキストに現れるキーワードほど値が小さくなる重み）で求めるのが一般的である。また、頻度ｆ（ｔ，ｋ）は、単純にテキストｔ中のキーワードｋの出現回数としてもよいが、テキストｔのテキスト長によって正規化した値であってもよい。ｓ（ｕ，ｒ）を計算する時点でｄｕすなわちｕのウェブ文書が未取得である場合は、ｓｉｍ（ｄｕ，ｒ）の値は０とする。上記の式から分かるように、ｄｕが未取得であっても、ｕが収集要求ｒを満足する可能性の大小が、ｕをリンクするｖのスコアや、そのリンクのアンカーテキストに基づいて推測できる。このようにして個々の収集要求ｒに対するｕのスコアｓ（ｕ，ｒ）が求められるが、Ｒ中の全ての収集要求ｒについてのｓ（ｕ，ｒ）の最大値をｓ（ｕ，Ｒ）とする。すなわち、

ｓ（ｕ，Ｒ）＝Ｍａｘ｛ｓ（ｕ，ｒ）｝（ここで、ｒ∈Ｒ）

である。ｓ（ｕ，Ｒ）の値が大きいｕほど、全てのＲを考慮した上で最も優先的に収集すべきＵＲＬであるとみなすことができる。

【００５０】

ｓ（ｕ，ｒ）とｓ（ｕ，Ｒ）の計算方法は、上記に説明した方法に限らない。ウェブ文書が未取得のＵＲＬに対して、取得する優先順位が十分に精度良く決定できる計算方法であれば、他の計算方法を採用してもよい。優先順位の精度がよいほど、ウェブ文書を取得するコストに対して、収集要求を満たす情報が収集できる割合が高くなる。ｓ（ｕ，ｒ）とｓ（ｕ，Ｒ）は、図１４におけるステップ１４３とステップ１４１４のように、新たなＵＲＬに対して常に計算される。また、既知のＵＲＬに対しても、ステップ１４５とステップ１４１２のように、Ｒの内容が変更される毎、ｕのウェブ文書やｕのリンク元のスコアが変化する毎にも計算される。図１４のステップ１４４で、ある収集要求ｒのキーワードの条件が変更された場合には、ステップ１４５にて、ｓ（ｕ，ｒ）とｓ（ｕ，Ｒ）が計算し直される。

【００５１】

ｓ（ｕ，ｒ）とｓ（ｕ，Ｒ）をつねに最新の値に維持した上で、ステップ１４６では、ＵＲＬ集合Ｕの中から、ウェブ文書をまだ取得していないｕを選択するか、もしくは、最後にウェブ文書を取得してから閾値以上の時間が経過したＵＲＬで、かつ、スコアｓ（ｕ，Ｒ）が最大であるようなｕを選択する。そこで、ｕが存在すれば（ステップ１４７）、このｕが、情報ネットワークから最優先に取得すべきＵＲＬである。ステップ１４７において、ｕが一つも存在しなければ、取得すべきＵＲＬがないので、処理を終了する（ステップ１４８）か、もしくは、収集要求集合Ｒの変更の有無をチェックしつつ処理を待機することになる。ステップ１４９では、ｕのウェブ文書を取得する。本実施形態が対象とするインターネットのウェブ文書については、ＨＴＴＰプロトコルに従った取得を行う。取得に失敗すれば（ステップ１４１０）、前のステップに戻り、他のＵＲＬに対して上述の処理を繰り返し行う。取得に成功すれば、これを図１のウェブ文書記憶部７に記憶する（ステップ１４１１）。次に、ｕのウェブ文書の内容に基づいて、上述のｓｉｍ（ｄｕ，ｒ）の項を計算して、スコアｓ（ｕ，ｒ）およびｓ（ｕ，Ｒ）を計算し直す（ステップ１４１２）。その後、取得したウェブ文書のパージング（タグの解析）を行って、当該ウェブ文

書がリンクするリンク先ＵＲＬを抽出し、その各々のｖについて（ステップ１４１３）、スコアｓ（ｖ，ｒ）およびｓ（ｖ，Ｒ）を計算し、ＵＲＬ集合Ｕにｖを追加する（ステップ１４１４）。情報収集部６は、以上に説明した処理を再帰的に行い、複数のコミュニティの全ての収集要求に対して、一括して並列に、要求を満たす可能性の高いウェブ文書を収集する。したがって、個々の収集要求毎に独立にクローリングを行って収集する場合と比べて、不要なウェブ文書を取得する割合が減るとともに、一つのトピックに着目したクローリングでは発見しにくいような、新たな情報を発見する機会が増えるという効果がある。

【００５２】

図１４のステップ１４５、ステップ１４１２、及びステップ１４１４でスコアを計算したＵＲＬのうち、ウェブ文書を取得済みのＵＲＬの中には、個々のコミュニティの収集結果として追加すべきものがある。あるいは逆に、収集結果の中にすでに含まれているＵＲＬのうち、収集要求の条件を満たさなくなったＵＲＬについては、これを収集結果から削除する必要がある。そこで、収集結果生成部８が行う処理を図１５を参照して説明する。

【００５３】

まず、対象とするｕのウェブ文書が取得済みであれば（ステップ１５１）、収集要求集合Ｒの中の、スコアｓ（ｕ，ｒ）が変化した収集要求について、下記の処理を繰り返し行う（ステップ１５２）。すなわち、収集要求ｒに対応する収集結果ｃに既にｕが含まれていれば（ステップ１５３）、収集要求ｒの各々のトピックにキーワードの論理式の形式で記述された条件をｕが満たすかどうかを調べる（ステップ１５４）。この処理は、ｕのウェブ文書のテキストが、収集要求ｒの論理式を満足する形でキーワードを含むかどうかを調べることによってなされる。ｕのウェブ文書のテキストが、収集要求ｒの中のどのトピックの条件も満たさなければ、ｕを収集結果ｃから削除する必要がある。しかし、過去にユーザがｕを有用であるとみなし、収集結果ｃの中にｕを含めるように明示的に編集を行ったことがある場合には（ステップ１５５）、ｕは収集結果ｃから削除しない。ステップ１５５において、明示的な編集とは、前述の図１３で示したような編集手段を用いて、ｕを追加したり、あるいはコメント文などの付加情報の作成を行う編集をいう。ステップ１５５において、ユーザが明示的な編集を行っていない場合は、ｕを収集結果ｃから削除する（ステップ１５６）。一方、ステップ１５３にて、ｕが収集結果ｃに含まれておらず、かつ、ｕが収集要求ｒの条件を満たす（ステップ１５７）ならば、ｕは収集結果ｃに追加すべきである。ただし、過去にユーザがｕを不要であるとみなし、収集結果ｃの中にｕを含めないように明示的に編集を行ったことがある場合には（ステップ１５８）、ｕを収集結果ｃに追加しない。ステップ１５８において、明示的な編集とは、前述の図１３で示したような編集手段を用いてｕを削除した場合をいう。このような場合以外は、ｕを収集結果ｃに追加する（ステップ１５９）。ここで、本実施形態の収集結果は、図１２と図１３で説明したように、トピックとサイトによって整理した形式で作成されるので、ｕを収集結果ｃの中のトピックのうち、条件を最もよく満たすトピックの中に追加する。また、ｕが既知のサイト内のＵＲＬである場合には、そのサイトの詳細情報として、図１２の情報１２５に示したような形で追加するし、未知のサイトの情報である場合には、図１２の情報１２８に示したように新しいサイトとして追加し、コメント文１２９としてウェブ文書から取得したテキストを付加する。

【００５４】

本発明の実施形態に係る情報収集システムにおいては、収集要求と収集結果を、ユーザが明示的に編集するだけなく、コミュニティ内でやり取りしたメッセージから収集要求と収集結果を自動的に更新する処理をも行う。この処理によって、動的に変化するユーザの興味・関心に常に合致するように収集要求と収集結果とを維持することができる。

【００５５】

図１６を用いて、メッセージに基づいて収集要求と収集結果を更新する処理の流れを説明する。

【００５６】

未処理のメッセージｍについて（ステップ１６１）、まず、ｍの返信メッセージを再帰的に集め、ｍを含むこれらのメッセージの集合をＭｍとする（ステップ１６２）。図１７に示したメッセージの例では、メッセージ１７１に対して、メッセージ１７２、１７３等が返信メッセージである。次に、Ｍｍのメッセージの各々から、ＵＲＬの記述、すなわち、「ｈｔｔｐ：／／」等で始まる記述を抽出して、これをＭｍ全てのメッセージについて集めたＵＲＬ集合をＵｍとする（ステップ１６３）。図１７の例では、１７４、１７６、１７８、１７１２がＵＲＬである。なお、テキスト１７１１は、ＵＲＬ１７４と同一であるし、メッセージ１７１の引用部分に含まれるので、この部分は処理しない。ステップ１６３の処理と同時に、Ｕｍの各ＵＲＬに対してメッセージ中に記述されているコメント文を抽出し、Ｕｍの各要素に対応したコメント文集合Ｄｍを得る（ステップ１６４）。ステップ１６４において、メッセージからＵＲＬへのコメント文を抽出する処理は、単純には、ＵＲＬと同一メッセージ内の同一の段落のテキストをそのまま抽出することで実現できるが、より複雑には、メッセージの返信関係に基づき、引用されているテキストまでも含めて文脈を理解し、複数のメッセージ間にまたがってコメント文を抽出する方法もある。図１７の例では、ＵＲＬ１７４に対するテキスト１７５、ＵＲＬ１７６に対するテキスト１７７、ＵＲＬ１７８に対するテキスト１７９、および、ＵＲＬ１７１２に対するテキスト１７１１が、コメント文として抽出される。また、ＵＲＬ１７１２はＵＲＬ１７１０（すなわち１７４）のサイト内のＵＲＬであり、さらに、ＵＲＬ１７１０はメッセージ１７１を引用した部分に含まれることから、テキスト１７１１およびＵＲＬ１７１２は、ＵＲＬ１７４をより詳細に説明する情報であると解釈できる。

【００５７】

このようにして、ＵＲＬ集合Ｕｍとコメント文集合Ｄｍとをメッセージ集合Ｍｍから得た後は、これを当該コミュニティの収集要求ｒ（または収集結果ｃ）の、どのトピックに追加すべきかを決定する処理を行う。

【００５８】

まず、ステップ１６５にて、収集要求ｒの各トピックに記述された収集起点ＵＲＬ（あるいは、収集結果ｃの各トピックに記述されたＵＲＬ）と、前記Ｕｍとを比較し、最も重複の多いトピックｔｍを選択することを試みる（ステップ１６５）。ＵＲＬの重複を調べる処理では、ＵＲＬが完全に一致する場合だけでなく、ＵＲＬのサイトが一致する場合も考慮する。ステップ１６５でｔｍが選択できない場合（ステップ１６６）には、収集要求ｒの各トピックに記述されたキーワード集合（あるいは収集結果ｃの各トピックに記述されたサイト名やコメント文などのテキスト）と、Ｄｍのテキストとを比較し、最も重複の多いトピックをｔｍとする（ステップ１６７）。ステップ１６７でもｔｍが選択できない場合（ステップ１６８）には、トピックを新たに作成してこれをｔｍとする（ステップ１６９）。この場合、トピック名には、メッセージのタイトルを用いる。さらに、収集要求を更新する場合には、新規トピックであるｔｍに対するキーワードとして、Ｄｍから抽出した重要語を選択する（ステップ１６１０）。ここでの重要語は、コメント文テキストに高い頻度で含まれ、かつ、他のトピックのコメント文テキストには低い頻度でしか含まれない語とする（従来の統計的手法により求めることができる）。ステップ１６５から１６１０の処理でトピックｔｍを選択もしくは作成した後、ｔｍに、先のＵｍを（収集結果の更新の場合には、Ｄｍのコメント文と関連付けて）追加する（ステップ１６１１）。

【００５９】

以上に説明した処理によって、図１７のメッセージに対して、図１８に示した収集要求、および、図１９に示した収集結果が生成される。図１８のトピック名１８１は図１７のメッセージ１７１のタイトルであり、キーワード１８２は、図１７のテキスト１７５、１７７、１７９、１７１１から抽出した重要語のＯＲからなる論理式である。また、収集起点ＵＲＬ１８３には、ＵＲＬ１７４、１７６、１７８、１７１２が設定される。ユーザは、自動的に生成されたこれらの項目を、必要ならば前述の収集要求編集手段を用いて適宜修正して、メッセージで議論された話題に関連する情報を収集するための収集要求を簡単に作成することができる。一方、図１９の収集結果については、トピック名１９１には図１

７のメッセージ１７１のタイトルが用いられ、サイト１９２、１９５、１９７にはそれぞれ図１７のＵＲＬ１７４、１７６、１７８が用いられる。各サイトに対するコメント文１９３、１９６、１９８には、それぞれ、図１７のテキスト１７５、１７７、１７９が用いられる。また、メッセージ１７３の１７１１の部分は、サイト１９２の詳細情報として情報１９４に示した形で埋め込まれる。このようにして自動生成された収集結果は、常にユーザにとって利用しやすい内容に作られるとは限らず、例えばコメント文１９８のように余分なテキストが含まれる場合もある。この場合には、前述の収集結果編集手段を用いて、ユーザが見やすい形に自由に編集することが容易に行える。

【００６０】

以上に説明した処理によって、一連のメッセージＭｍに対して、収集要求あるいは収集結果のトピックｔｍが関連付けられる（ステップ１６５、１６７）か、あるいは、新たに作成される（ステップ１６９）。このようなメッセージとトピックとの関連をユーザに提示することによって、ユーザがメッセージを理解したり、メッセージと関連する情報にアクセスしたりする作業を支援することができる。これは例えば、図９に示したように、メッセージ「音楽配信ビジネス」９２に対して、関連するトピック「コンテンツ配信」９３を関連付けて表示することによって行われる。

【００６１】

一方、収集結果に対してユーザが行う編集に応じて、収集要求を自動的に更新することも可能である。この処理は、図１６で説明した処理と同様の処理で実現される。ユーザが自由な形式で記述するメッセージと異なり、収集結果は、上述の収集結果編集手段（図１３）で説明したような所定の形式で記述するため、この処理は図１６の処理よりも比較的容易に実現できる。収集要求の条件とするキーワードは、収集結果に記述されるコメント文等から作成する。

【００６２】

図１のウェブ文書検索部１０の処理の流れを、図２０を用いて説明する。ウェブ文書検索部１０は、図１の情報収集部６が収集してウェブ文書記憶部７に記憶したウェブ文書を、ユーザが検索して利用するための手段である。

【００６３】

図２０において、まず、ユーザによって検索条件ｑが入力されると（ステップ２０１）、収集済みのウェブ文書からｑを満足する文書を検索し、その結果のＵＲＬ集合をＵｑとする（ステップ２０２）。次に、Ｕｑの各々の要素ｕについて（ステップ２０３）、ｕを含む収集結果ｃを探す（ステップ２０４）。この収集結果ｃは、ｕ自体を含む収集結果であってもよいし、あるいは、ｕと同一サイトのＵＲＬや、ｕをリンクするリンク元のＵＲＬを含む収集結果であってもよい。このような収集結果ｃが存在すれば（ステップ２０５）、ｕを説明する見出しおよび説明文として収集結果ｃに記述されているサイト名、コメント文のテキストを用い、ｕと収集結果ｃとを関連付けてユーザに提示する（ステップ２０６）。収集結果ｃが存在しなければ、ｕを説明する見出しおよび説明文として、ｕのウェブ文書に記述されているタイトルや本文等のテキストを用いてｕをユーザに提示する（ステップ２０７）。

【００６４】

図２１は、図２０で説明した処理によってユーザに提示された検索結果の画面例を示す図である。ユーザが入力した検索条件「オークション」２１１に対して検索された個々のウェブ文書のＵＲＬ「ｈｔｔｐ：／／ｘｙｚ．ｃｏｍ／」２１２等に対して、見出し「○○オークション」２１３、説明文２１４等を、ステップ２０４で求めた収集結果、例えば図１９に示すサイト名１９２、コメント文１９３を用いてユーザに提示する。さらに、図２１に示すように、収集結果のトピック２１５を収集結果と関連付けて提示する。検索結果のＵＲＬと関連する収集結果がなければ、例えば、検索結果の説明文としてウェブ文書のテキストの一部２１７（一般的には、冒頭部分のテキストや、検索語が出現する近傍のテキスト）を提示する。このように、ウェブ文書からそのまま得たテキストは、意味が理解し難しかったり、必ずしもそのサイトの内容を適切に表した記述でない場合がある。これ

に対し、説明文２１４のように、コミュニティのメンバーが収集結果の中で記述したテキストは、簡潔で理解しやすい記述である場合が多い。また、検索結果の情報に対して図２１に示すように収集結果のトピックを関連付けて表示することにより、その情報がどのような分野・文脈の情報であるかが容易に理解できるようになる。さらに、ユーザは、当該トピックに含まれる他の有用な情報を利用することができる。あるトピックに関する情報を収集しているコミュニティは、そのトピックに関心を持つ専門家の集団であると言えるので、検索結果中の個々の情報について、どのようなコミュニティがこれを有用とみなしているか、いないかを、即座に知ることができるという効果もある。

【００６５】

以上に説明した処理は、検索結果と収集結果とを関連付けて提示する処理であったが、これと同様の方法により、あるコミュニティの収集結果に対して、他のコミュニティの収集結果を関連付けて表示することも可能である。

【００６６】

図１２の情報１２７の例では、「ｅコマース調査グループ」が「コンテンツ配信」のトピックとして収集した情報「××エンターテインメント」に対し、別のコミュニティである「カラオケ友の会」が収集した「家庭用コンテンツ」のトピック１２７が関連付けて提示される。この処理も、図２０のステップ２０４と同様に、あるＵＲＬが収集結果に含まれているかどうかを調べることって実現される。このように、検索結果や収集結果に対し、他のコミュニティが関心のあるトピックや収集した情報を関連付けて提示することは、ユーザが検索結果や収集結果を利用する際の手助けになるだけでなく、ユーザが参加していない他のコミュニティがどのようなトピックに関心を持って活動を行っているかを、知る機会を増やす働きをする。その結果、複数のコミュニティ間の交流が活発になるという効果がある。

【００６７】

本発明は、上記の発明の実施の形態に限定されるものではない。本発明の要旨を変更しない範囲で種々変形して実施できるのは勿論である。

【００６８】

【発明の効果】

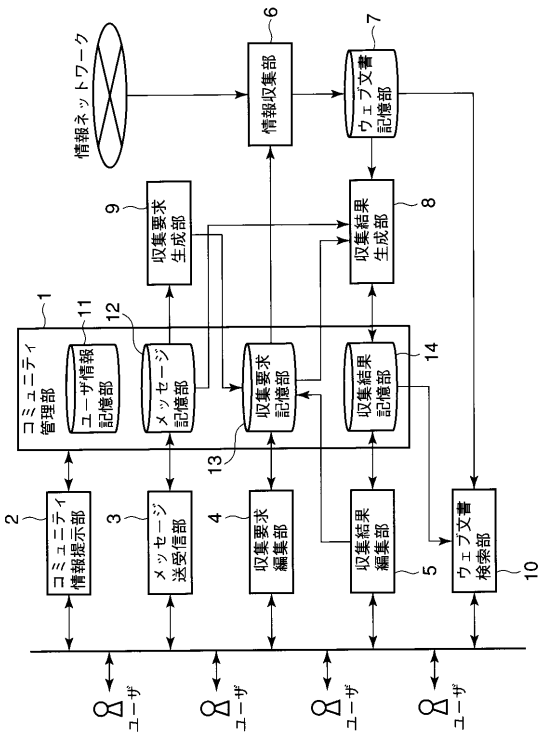
以上説明したように、本発明によれば、共通の関心を持ったコミュニティのメンバーが共同で収集要求と収集結果を編集し、これを継続的に洗練・保守していくことができるので、メンバー一人一人の少ない労力の寄与によって、コミュニティ全員にとって有用な情報を収集・整理して共有することができる。さらに、コミュニティ内で日常的に行われるメッセージのやり取りに基づいて、収集要求と収集結果が自動的に更新されるので、収集要求と収集結果を編集するユーザの作業が軽減するとともに、コミュニティの活動に応じて動的に変化する関心に対応した情報収集を行うことができる。

【図面の簡単な説明】

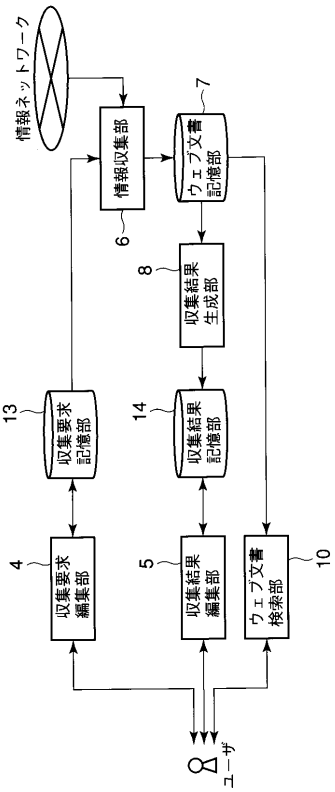
【図１】　本発明の一実施形態である情報収集システムの構成を示す図。
【図２】　従来の情報収集システムの構成の一例を表す図。
【図３】　ユーザ情報の例を表す図。
【図４】　ユーザの登録、認証およびコミュニティへの参加の処理の流れを表す図。
【図５】　コミュニティ情報の一覧提示画面の例を表す図。
【図６】　メッセージの送信の処理の流れを表す図。
【図７】　収集要求の編集の処理の流れを表す図。
【図８】　収集結果の編集の処理の流れを表す図。
【図９】　メッセージの閲覧画面の例を表す図。
【図１０】　メッセージの編集画面の例を表す図。
【図１１】　収集要求の編集画面の例を表す図。
【図１２】　収集結果の閲覧画面の例を表す図。
【図１３】　収集結果の編集画面の例を表す図。
【図１４】　情報収集の処理の流れを表す図。

【図１５】　収集結果の生成の処理の流れを表す図。
【図１６】　メッセージから収集要求または収集結果を生成する処理の流れを表す図。
【図１７】　メッセージの例を表す図。
【図１８】　メッセージから生成された収集要求の例を表す図。
【図１９】　メッセージから生成された収集結果の例を表す図。
【図２０】　ウェブページ検索の処理の流れを表す図。
【図２１】　ウェブページ検索の検索結果画面の例を表す図。
【符号の説明】
１…コミュニティ管理部
２…コミュニティ情報提示部
３…メッセージ送受信部
４…収集要求編集部
５…収集結果編集部
６…情報収集部
７…ウェブ文書記憶部
８…収集結果生成部
９…収集要求生成部
１０…ウェブ文書検索部
１１…ユーザ情報記憶部
１２…メッセージ記憶部
１３…収集要求記憶部
１４…収集結果記憶部

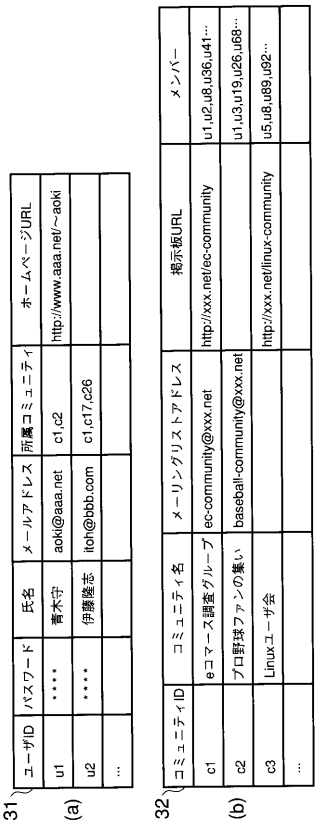
【図１】



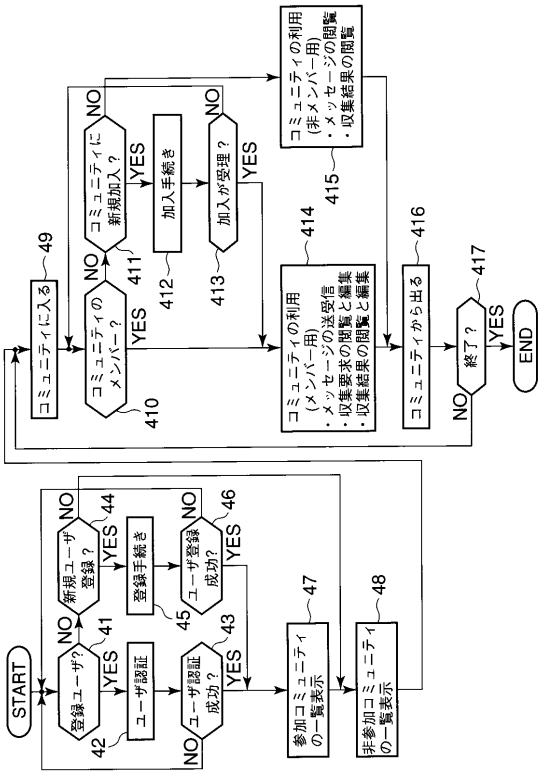
【図２】



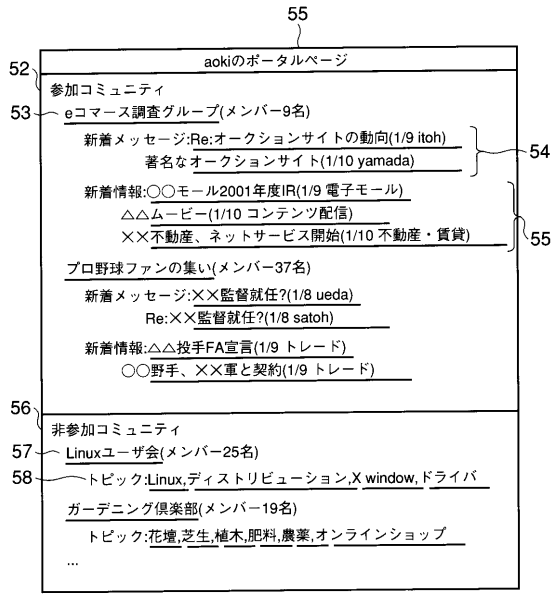
【図３】


ユーザID
パスワード
氏名
メールアドレス
所属コミュニティ
ホームページURL
u1
****
青木守
aoki@aaa.net
c1,c2
http://www.aaa.net/~aoki
u2
****
伊藤隆志
itoh@bbb.com
c1,c17,c26
...
31
(a)
コミュニティID
コミュニティ名
メーリングリストアドレス
掲示板URL
メンバー
c1
eコマース調査グループ
ec-community@xxx.net
http://xxx.net/ec-community
u1,u2,u8,u36,u41...
c2
プロ野球ファンの集い
baseball-community@xxx.net
u1,u3,u19,u26,u68...
c3
Linuxユーザ会
http://xxx.net/linux-community
u5,u6,u89,u92...
...
32
(b)

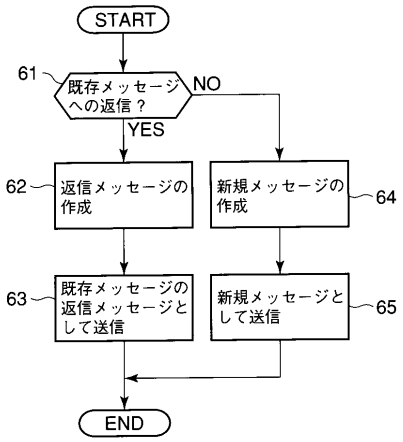
【図４】


START
登録ユーザ？
NO
YES
41
新規ユーザ登録？
44
登録手続き
45
ユーザ登録成功？
46
ユーザ認証
42
ユーザ認証成功？
43
参加コミュニティの一覧表示
47
非参加コミュニティの一覧表示
48
コミュニティに入る
49
コミュニティのメンバー？
410
コミュニティに新規加入？
411
加入手続き
412
加入が受理？
413
コミュニティの利用（メンバー用）
・メッセージの送受信
・収集要求の閲覧と編集
・収集結果の閲覧と編集
414
コミュニティの利用（非メンバー用）
・メッセージの閲覧
・収集結果の閲覧
415
コミュニティから出る
416
終了？
417
END

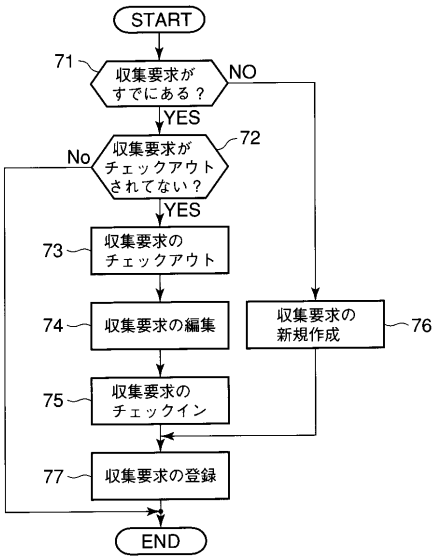
【図５】


55
aokiのポータルページ
52
参加コミュニティ
53
eコマース調査グループ(メンバー9名)
新着メッセージ:Re:オークションサイトの動向(1/9 itoh)
著名なオークションサイト(1/10 yamada)
54
新着情報:○○モール2001年度IR(1/9 電子モール)
△△ムービー(1/10 コンテンツ配信)
××不動産、ネットサービス開始(1/10 不動産・賃貸)
55
プロ野球ファンの集い(メンバー37名)
新着メッセージ:××監督就任?(1/8 ueda)
Re:××監督就任?(1/8 satoh)
新着情報:△△投手FA宣言(1/9 トレード)
○○野手、××軍と契約(1/9 トレード)
56
非参加コミュニティ
57
Linuxユーザ会(メンバー25名)
58
トピック:Linux,ディストリビューション,X window,ドライバ
ガーデニング倶楽部(メンバー19名)
トピック:花壇,芝生,植木,肥料,農薬,オンラインショップ
...

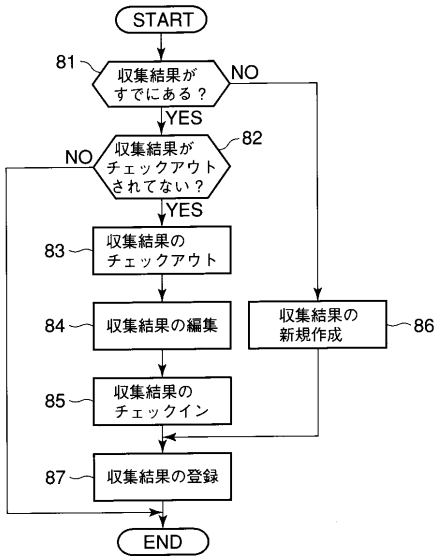
【図６】


START
61
既存メッセージへの返信？
NO
YES
62
返信メッセージの作成
新規メッセージの作成
64
63
既存メッセージの返信メッセージとして送信
新規メッセージとして送信
65
END

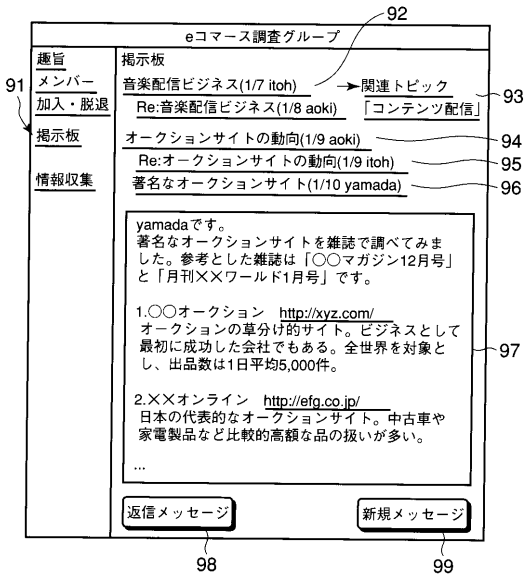
【図７】


START
71
収集要求が
すでにある？
NO
YES
No
72
収集要求が
チェックアウト
されてない？
YES
73
収集要求の
チェックアウト
74
収集要求の編集
収集要求の
新規作成
76
75
収集要求の
チェックイン
77
収集要求の登録
END

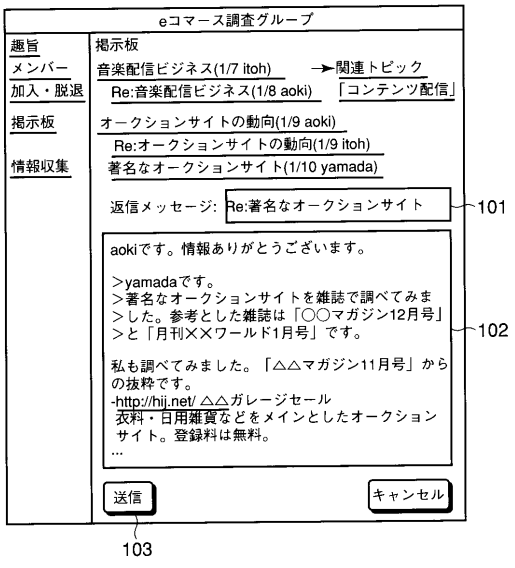
【図８】


START
81
収集結果が
すでにある？
NO
YES
NO
82
収集結果が
チェックアウト
されてない？
YES
83
収集結果の
チェックアウト
84
収集結果の編集
収集結果の
新規作成
86
85
収集結果の
チェックイン
87
収集結果の登録
END

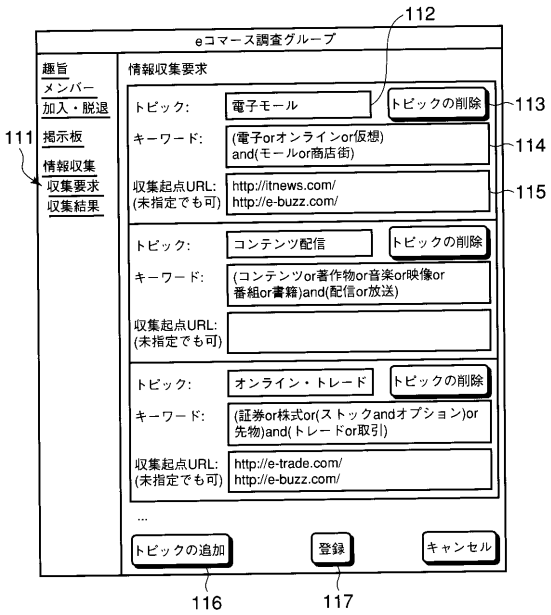
【図９】


92
eコマース調査グループ
趣旨
メンバー
加入・脱退
掲示板
情報収集
91
掲示板
音楽配信ビジネス(1/7 itoh)
→関連トピック
Re:音楽配信ビジネス(1/8 aoki)
「コンテンツ配信」
93
オークションサイトの動向(1/9 aoki)
94
Re:オークションサイトの動向(1/9 itoh)
95
著名なオークションサイト(1/10 yamada)
96
yamadaです。
著名なオークションサイトを雑誌で調べてみました。参考とした雑誌は「○○マガジン12月号」と「月刊××ワールド1月号」です。
1.○○オークション http://xyz.com/
オークションの草分け的サイト。ビジネスとして最初に成功した会社でもある。全世界を対象とし、出品数は1日平均5,000件。
2.××オンライン http://efg.co.jp/
日本の代表的なオークションサイト。中古車や家電製品など比較的高額な品の扱いが多い。
...
97
返信メッセージ
新規メッセージ
98
99

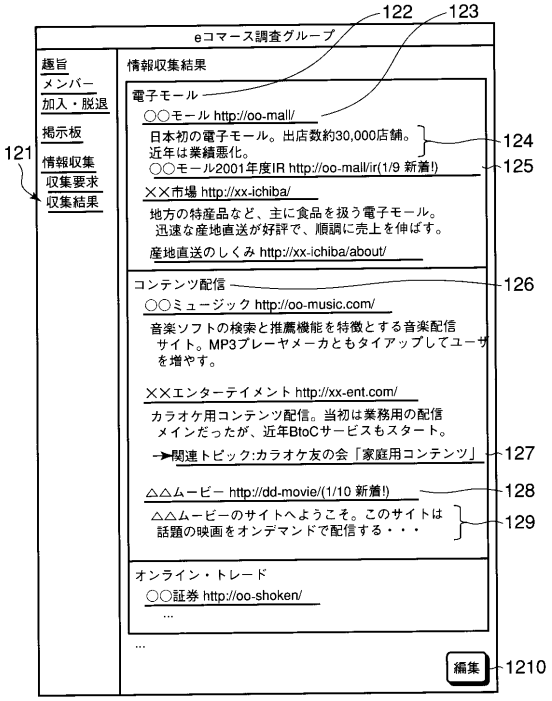
【図１０】


eコマース調査グループ
趣旨
メンバー
加入・脱退
掲示板
情報収集
掲示板
音楽配信ビジネス(1/7 itoh)
→関連トピック
Re:音楽配信ビジネス(1/8 aoki)
「コンテンツ配信」
オークションサイトの動向(1/9 aoki)
Re:オークションサイトの動向(1/9 itoh)
著名なオークションサイト(1/10 yamada)
返信メッセージ: Re:著名なオークションサイト
101
aokiです。情報ありがとうございます。
>yamadaです。
>著名なオークションサイトを雑誌で調べてみま
>した。参考とした雑誌は「○○マガジン12月号」
>と「月刊××ワールド1月号」です。
私も調べてみました。「△△マガジン11月号」からの抜粋です。
-http://hij.net/ △△ガレージセール
衣料・日用雑貨などをメインとしたオークションサイト。登録料は無料。
...
102
送信
キャンセル
103

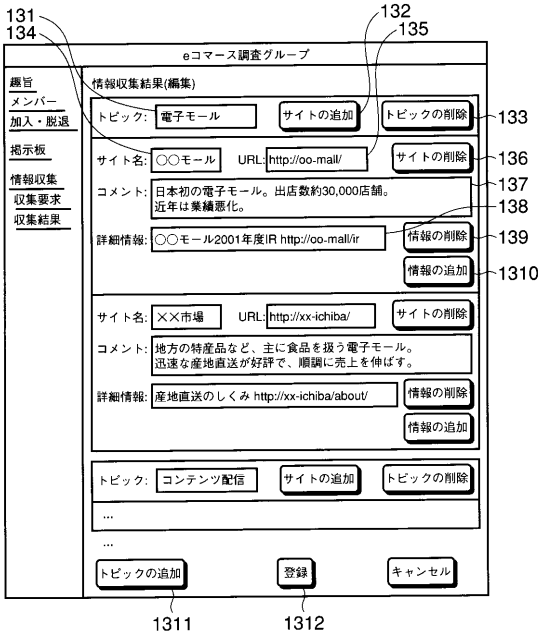
【図１１】



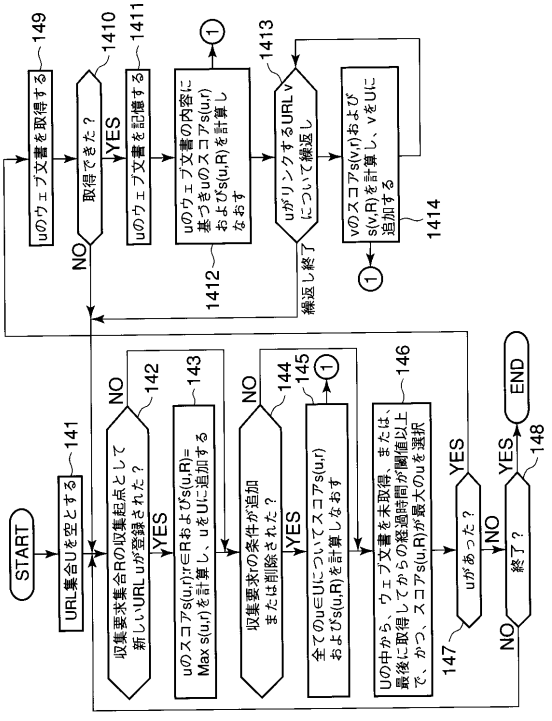
【図１２】



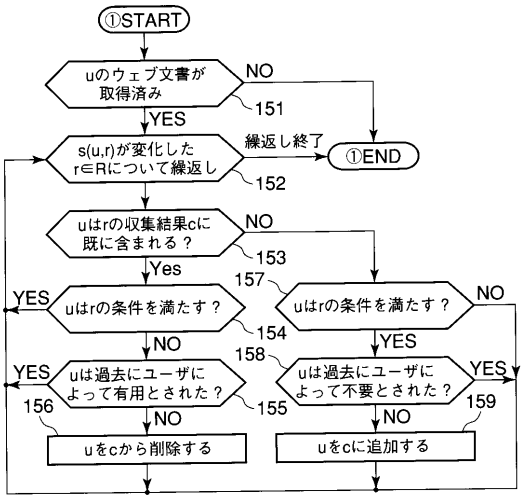
【図１３】



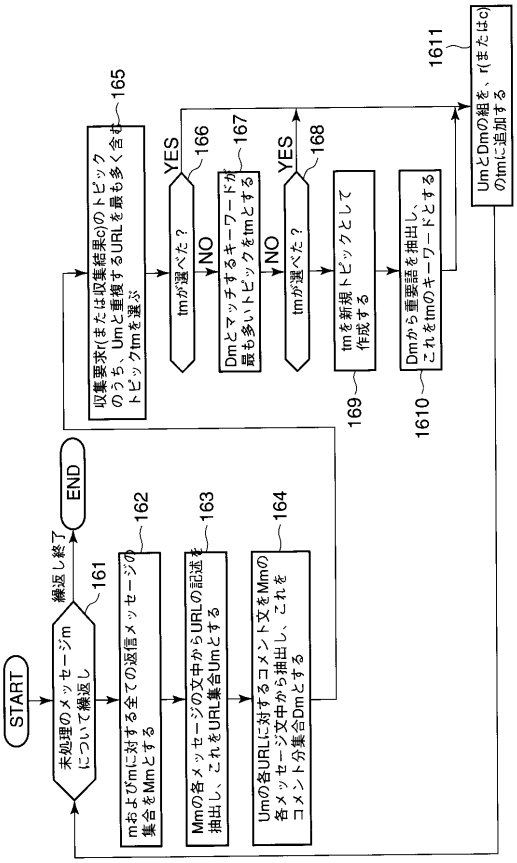
【図１４】



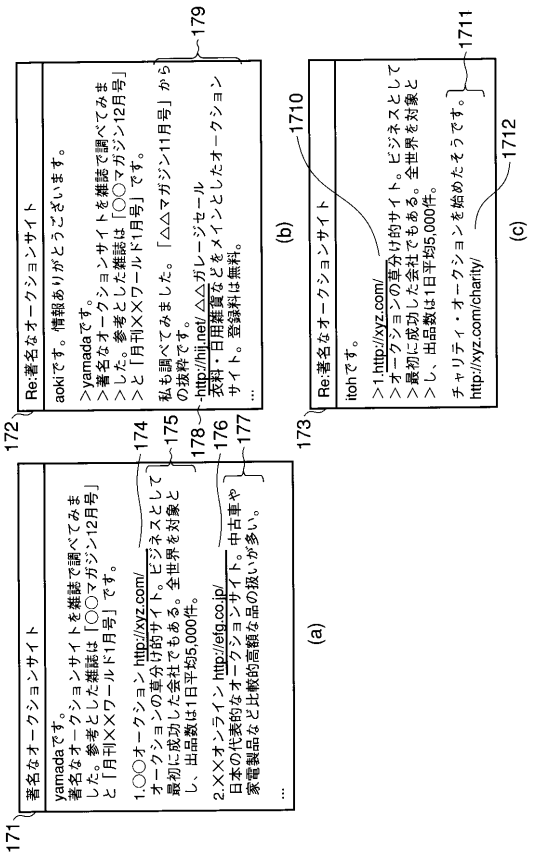
【図１５】



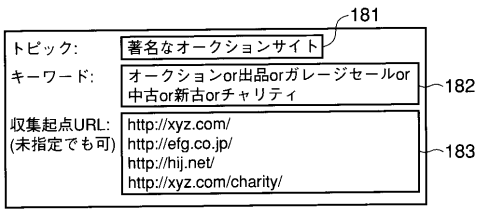
【図１６】



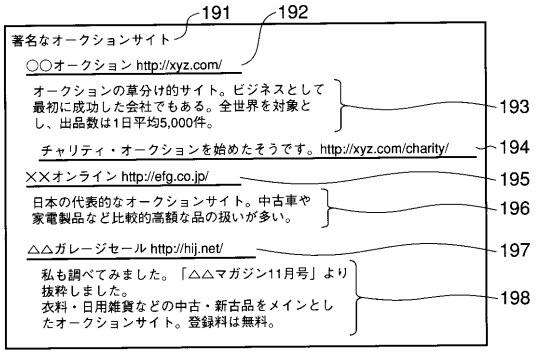
【図１７】



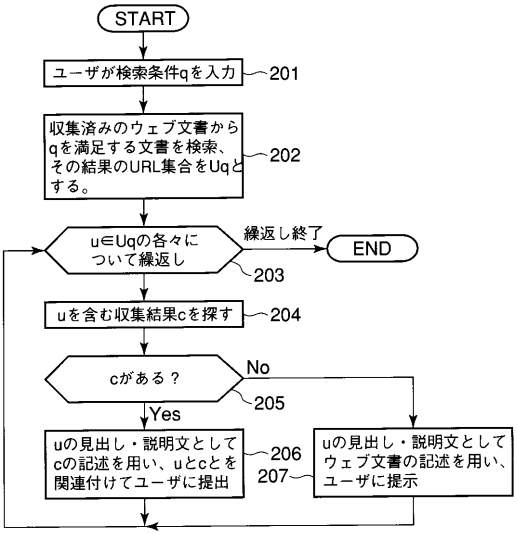
【図１８】



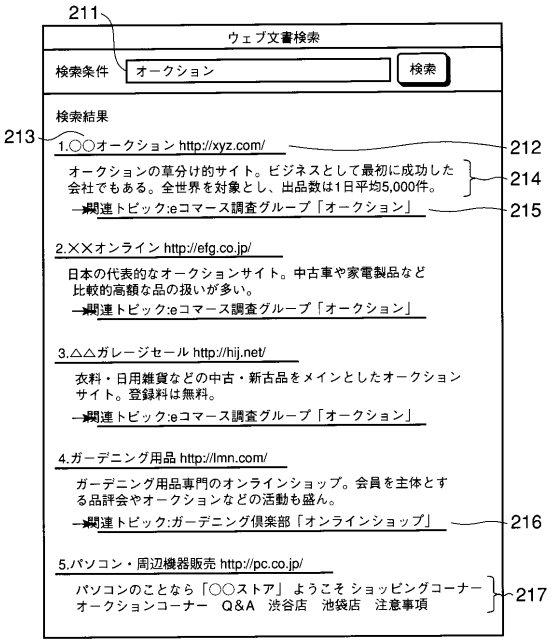
【図１９】



【図２０】


START
ユーザが検索条件qを入力
201
収集済みのウェブ文書からqを満足する文書を検索、その結果のURL集合をUqとする。
202
u∈Uqの各々について繰返し
繰返し終了
END
203
uを含む収集結果cを探す
204
cがある？
No
205
Yes
uの見出し・説明文としてcの記述を用い、uとcとを関連付けてユーザに提出
206
207
uの見出し・説明文としてウェブ文書の記述を用い、ユーザに提示

【図２１】


211
ウェブ文書検索
検索条件
オークション
検索
検索結果
213
1.○○オークション http://xyz.com/
212
オークションの草分け的サイト。ビジネスとして最初に成功した会社でもある。全世界を対象とし、出品数は1日平均5,000件。
214
→関連トピック:eコマース調査グループ「オークション」
215
2.××オンライン http://efg.co.jp/
日本の代表的なオークションサイト。中古車や家電製品など比較的高額な品の扱いが多い。
→関連トピック:eコマース調査グループ「オークション」
3.△△ガレージセール http://hij.net/
衣料・日用雑貨などの中古・新古品をメインとしたオークションサイト。登録料は無料。
→関連トピック:eコマース調査グループ「オークション」
4.ガーデニング用品 http://lmn.com/
ガーデニング用品専門のオンラインショップ。会員を主体とする品評会やオークションなどの活動も盛ん。
→関連トピック:ガーデニング倶楽部「オンラインショップ」
216
5.パソコン・周辺機器販売 http://pc.co.jp/
パソコンのことなら「○○ストア」 ようこそ ショッピングコーナー
オークションコーナー Q&A 渋谷店 池袋店 注意事項
217

フロントページの続き

(72)発明者　後藤　和之
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

審査官　梅本　達雄

(56)参考文献　特開２００１－３１２５１５（ＪＰ，Ａ）
特開２００１－３４４２５７（ＪＰ，Ａ）
特開２００１－１３４６１６（ＪＰ，Ａ）
イントラネット対応が進む全文検索システム，日経コンピュータ　ｎｏ．４１４　NIKKEI COMPUTER，１９９７年　3月３１日

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30
G06F 13/00
JSTPlus(JDream2)