L_014

P2P ネットワークを基盤とした分散 Web 検索システムにおける索引情報の損失防止手法の検討

A Study on a Preventing Index-Loss Method for a Distributed Web Retrieval System on a P2P Network

豊田 正隆　　勅使河原 可海
Masataka Toyoda　　Yoshimi Teshigawara

1．はじめに

近年，共通のジャンルを扱う個人サイトを作成して情報を公開している人々と，そのジャンルに興味を持つ閲覧者による，Web 上のコミュニティが増加している．それらのコミュニティに属する人々にとっては，コミュニティ内で発生する新しい情報を即座に検索することへの要求が大きい．

Web 上の情報を発見する手段としてサーチエンジンがある．しかし，一般のサーチエンジンは文書収集の対象範囲が広いため，対象範囲と実行間隔の間のトレードオフの関係から，上記の要求を満たすことができない．

文書収集の対象範囲を特定のコミュニティに絞ることで，文書収集の実行間隔を短くすることが可能になる．それによって要求を満たすことができると考えられるが，現状では特定のコミュニティのためだけのサーチエンジンは存在しない．また，そのためのサーチエンジンを新たに構築するには多大な費用と運用の手間がかかる．

そこで我々は，個人の所有する PC を利用した，個人サイト特化型分散 Web 検索システムの研究を行ってきた[1]．本稿では，システムの適切な動作と堅牢性を両立する索引情報の最適な複製数（重複度）を求め，PC の不足によって必要な複製を確保できない場合の動作についての検討を行ったので，これを報告する．

2．システムの概要

本システムは，30 分以前に Web 上で公開された情報を即座に検索することを目的とした分散検索システムであり，個人が提供する PC によって構成されるため安価に構築が可能であり，運用の手間を不要とする堅牢性を持つことを特徴としている．

本システムを構成している PC は，P2P ネットワークを構成し，ピアとして振舞う．各ピアは，あらかじめシステムに登録されたサイトの中から，自身が担当しているサイトに対して 30 分間隔でクローリングを行い，索引情報を作成することで，30 分以前に公開された情報の検索を可能としている．また，サイトをカテゴリに分け，検索時にカテゴリを指定することで，検索要求に対して一部のサーバのみが検索処理を行うようになっている．

ピアは図 1 のようにグループを構成し，同一グループに属するピアが同じ情報を保持することで，一部のピアの脱落による情報の損失を防ぐ．

システム上には，全てのピアから成るシステムグループ，1 つのカテゴリを担当するピアから成るカテゴリグループ，同じサイト群を担当するピアから成るサイトグループの 3 種類のグループが存在する．同一サイトグループに属するピアは同一の索引情報を保持しており，1 つのサイトグループを構成するピアの数が重複度となる（図 1 では重複度は 2）．文書収集は，サイトグループのピアの 1 つが代表して行う．システムグループのピアは，各カテゴリグループのいくつかのピアの情報を保持する．また，カテゴリグループのピアは，同一カテゴリグループ上の各サイトグループのピアと，その担当するサイトの情報を保持する．このように，いくつかのピアの情報を常に保持しておくことで，ピアの脱落時に代理のピアを即座に発見することができる．

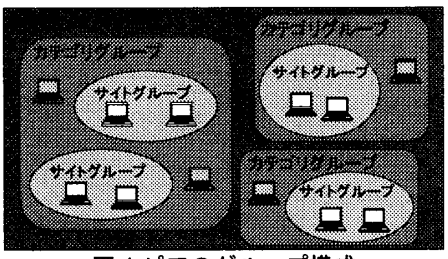

図 1 ピアのグループ構成

3. 最適な重複度の算出

以下の手順で最適な重複度を算出した．

(1) コミュニティにおける潜在的ピア数を仮定する
(2) ピアのオンライン率を調べる
(3) 求めたオンライン率と仮定した環境における最適な重複度を導く

3.1 ピア数の仮定

本システムの対象とするコミュニティの最小規模を次のように定めた．

・コミュニティを構成するサイトの数：50 サイト
・コミュニティに興味を持つ人数：500 人

コミュニティに興味を持つ人数はシステムのピアとなりうる潜在的ピア数であると考えられる．そのため，対象とするコミュニティの潜在的ピア数を 500 とする．

3.2 オンライン率の調査

本システムはオンラインになっている個人 PC を利用するため，個人のオンライン状態を示す Instant Messenger（以下 IM）のオンライン率を，本システムのピアの参加率・脱落率に適用できると考えた．ここで，参加率は「現在 P2P ネットワークに参加していないそのピアが，次の 1 分間で参加する確率」，脱落率は「P2P ネットワークに参加しているそのピアが，次の 1 分間で脱落する確率」を意味する．

50 名を対象に IM にサインインしている時間を調べた．結果を表 1 に示す．実験の結果から，ピアの参加率と脱落率を以下のように定めた．実際の環境では様々な要因によって参加率・脱落率が変化することが想定されるため，実験結果よりも悪い条件となるように設定している．

・参加率：0.05%　　・脱落率：1.00%

創価大学大学院工学研究科
Graduate School of Engineering, Soka University

表1 サインイン時間の調査結果

調査日	火曜日	土曜日
平均サインイン回数	2.30	1.38
平均サインイン時間	4:45:44	2:57:46
1分当たりのサインイン確率	0.20%	0.11%
1分当たりのサインアウト確率	0.80%	0.78%

3.3 最適な重複度の算出

3.1 と 3.2 で定めた潜在的ピア数と参加率・脱落率において，重複度を変化させた際に，サイトグループのピアが全て脱落する（サイトグループの崩壊）確率がどう変化するかをシミュレーションで調べた．以下にシミュレーションの内容を示す．

- システムは1つのサイトグループのみを持つ．
- システムが 24 時間（1440 分）動作した際の，サイトグループが崩壊する試行の割合を求める．
- 重複度は1から 10 の場合について調べる．
- 試行は各重複度の値につき 100 回行う．
- 初期状態では，25 のピアがオンラインであり，サイトグループのピア数は重複度に等しいとする．
- オフラインである全てのピアは，1 分ごとに参加率の確率で参加する．
- オンラインである全てのピアは，1 分ごとに脱落率の確率で脱落する．
- サイトグループのピアが脱落した場合，サイトグループに属していないオンラインのピアを無作為に選んでサイトグループに取り込む．その際，サイトグループのいずれかのピアが，10 分かけて取り込むピアへ索引情報を送信する．送信ピアが送信途中で脱落した場合，サイトグループの他のピアが送信をやり直す．索引情報を受信中のピアは，取り込むピアに索引情報を送信することはできない．また，取り込めるピアが存在しない場合は，新たなピアがオンラインになるまで待つ．

シミュレーションの結果を表2に示す．

表2 重複度と崩壊の発生率の関係

重複度	崩壊した試行の割合
1	1.00
2	0.91
3	0.45
4	0.08
5	0.01
6以上	0.00

重複度が 3 以下になると，崩壊した試行の割合が急激に増加していることが分かる．また，重複度が高いほどサイトグループが崩壊する確率が低いことが分かる．しかし，文書収集を行うピアが作成した索引情報を同一サイトグループに属する他のピアに送信する必要があるため，重複度を高くすると索引情報の転送を行う回数 s が大きくなる．索引情報を受け取ったピアは他のピアに索引情報を送信できるので，重複度を d[ピア]とすると，s は下の式で表される．

$$s = ceil(\log_2 d)$$

本システムは 30 分以内に更新された情報が検索可能であることを目標にしているため，30 分より短い時間で文書収集，索引情報の更新，他のアクティブピアへの送信を完了させる必要がある．先日行った実環境実験によると，1 つのサイトの文書収集および索引情報の更新にかかる時間の平均は約 1 分，索引情報のデータ量は約 1500kB となっている．また，ピアである個人 PC は ADSL によってインターネットに接続されている場合が多く，上り方向のデータ転送速度は約 640kbps である[1]（ここでは低く見積もって 400kbps とする）．

また，P2P ネットワーク上のピアの平均数 a[ピア]は，

$$a[\text{ピア}] = \text{潜在的ピア数} \times \frac{\text{参加率}}{\text{参加率} + \text{脱落率}} = 23.8$$

となり，サイトグループの数と重複度の積（全てのサイトグループが重複度を満たすために必要なピアの数）p[ピア]が a を上回ることはできない．以上から，1 つのサイトグループが担当できるサイトの数を c[サイト]，文書収集から索引情報の送信までにかかる時間を t[分]とすると，

$$\begin{cases} p[\text{ピア}] = \dfrac{50[\text{サイト}]}{c[\text{サイト}]} \times d[\text{ピア}] \\ t[\text{分}] = c[\text{サイト}] \times 1[\text{分}] + \dfrac{1500[kB] \times c[\text{サイト}]}{400[kbps] \div 8[bit] \times 60[\text{秒}]} \times s \end{cases}$$

となる（ceil(x)は小数点切り上げ関数）．制約条件として $p < 23.8$，$t < 30$ があるため，$d = 4$，$c = 10$ が最適となる（サイトグループの数は $50 / c = 5$）．

4. ピア不足時の動作

オンラインになっているピアの数は変動するため，ピアの不足によって 3.で定めた重複度分のピアを確保できないサイトグループが出てくる．サイトグループのピア数が 2 以下になった場合，表 2 から，その後のピアの脱落によってサイトグループが崩壊する可能性が高いことが分かる．そこで，サイトグループのピア数が 2 以下になった場合，ピア数が 4 であるサイトグループのピアを，そのサイトグループに「移籍」させることで，サイトグループが崩壊する確率を減少させる．サイトグループの数を 5 として，3.3 のシミュレーションと同じ条件の試行を 500 回シミュレーションした．結果を表 3 に示す．

表3 移籍の有無と崩壊の発生率の関係

移籍の有無	崩壊した試行の割合
移籍有り	0.59
移籍無し	0.68

表 3 から，ピアが補充できない状況であっても，移籍によってサイトグループの崩壊が生じる可能性を下げられることが分かる．

5. まとめと今後の課題

本稿では，P2P ネットワークを基盤とした分散 Web 検索システムにおける最適な重複度と，ピア不足時の動作についての検討を行い，その有効性をシミュレーションによって示した．

今後は，システムの実装と運用を行い，システムの有用性を検証していく．

参考文献

[1] 豊田正隆，山崎賢悟，勅使河原可海：個人の PC による P2P ネットワークを基盤とした柔軟な分散 Web 検索システムの提案，FIT2005 第 4 回情報科学技術フォーラム一般講演論文集第 4 分冊，pp.287-288，2005.9