Web コンテンツの自動選別システムの開発

03H022 樫尾豪人

1　はじめに

現在インターネットは世界中の人が利用しており、インターネット上には膨大な量の Web コンテンツがある。その中から、自分の興味のある Web コンテンツを新に捜し出すには大変な労力と時間が必要である。ポータルサイトや検索エンジンでは Web コンテンツを探しやすくするためのさまざまな機能があるが、一人々々の興味の対象から Web コンテンツを、自動的に選別する機能がない。また、大規模なポータルサイトや検索エンジンのような大規模ネットワークでは、利用者が多すぎる。そのため、興味の対象が分散しすぎ、実現しにくい機能である。小規模ネットワークとして考えた場合では、その中でインターネットを利用する人は同じグループに所属していたり、同じことを調べたりする可能性が高いと考えられる。そのことから、興味や調べる対象が似ている可能性が高いと思われる。そこで、小規模ネットワーク内から見た Web コンテンツのログと自分の興味のあるキーワードに対する Web コンテンツの URL とを比較する。そこで合致したものを抽出することにより、検索エンジンでキーワードを用いた検索を行うより、利用価値の高い Web コンテンツを優先的に見つけ出すシステムを開発した。

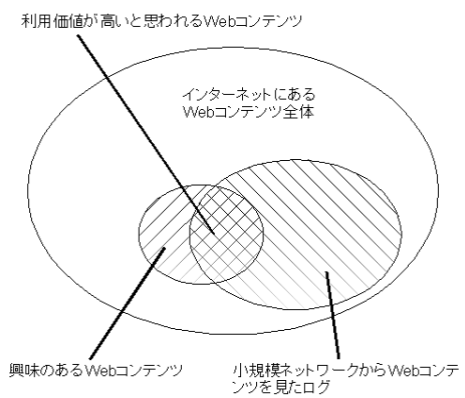

図 1　Web コンテンツの分布

2　Web コンテンツの自動選別システム

Web コンテンツの選別には、小規模ネットワーク内で閲覧された Web コンテンツに対しての URL のログ、興味のあるキーワードのデータベースを用いた。手順は以下の通りにした。

1. 興味のあるキーワードをデータベースに登録する
2. 興味のあるキーワードのデータベースを用いて、Google SOAP Search API を使い検索を行う
3. Google SOAP Search API から帰ってきた結果と小規模ネットワーク内から Web コンテンツを見たログとを比較する
4. 合致するものと、合致しなかったものを別々に出力する

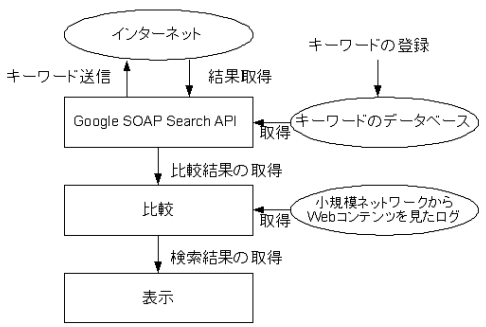

図 2　Web コンテンツの選別

3　まとめ

Web コンテンツ自動選別システムを開発した。それにより、より早く簡単に興味があるコンテンツを見つけ出すことができた。そのため、Web コンテンツを見つけ出すための時間や労力を、軽減することができるとされる。