

Rosette®文字・言語コード判別システム

文部科学省リーディングプロジェクト様

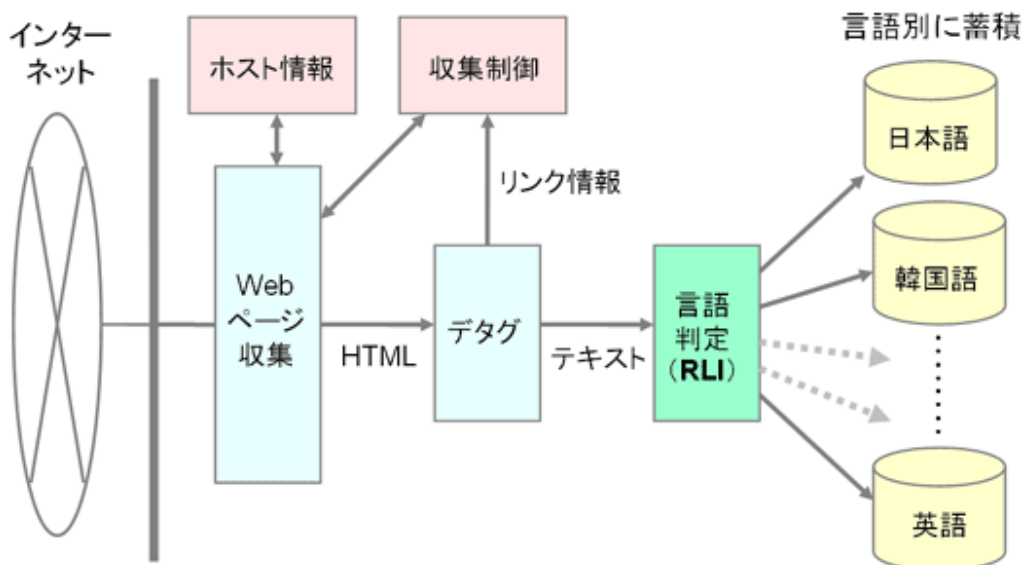
プロジェクト概要：

2003年度から実施している文部科学省リーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」の研究開発課題「インターネット上の知識集約を可能にするプラットフォーム構築技術」（研究代表者：早稲田大学理工学部 村岡洋一教授）では、Web 情報の大規模な収集ならびにこれらの情報を対象としたマイニング技術の開発がおこなわれています。2003年度には10億URLのデータ収集ならびにそれら Web データのリンク情報の解析がおこなわれ、2005年度までに約120億規模の Web データを収集し、必要情報抽出のための知識フィルタリング技術の開発に取り組みます。

Basis Technology 製品導入背景：

Web 上のデータは、言語および文字コード情報が不明なものや、文字コードが指定されていても間違っているものがあり、結果として文字化けを生じてしまい適切に利用することができなくなるものがあります。本プロジェクトでは、膨大な多言語情報の収集および解析をおこなうため、データの言語ならびに文字コードの正確な判別と文字化けをなくすことが大前提でした。そこで、Basis Technology の Rosette 言語・文字コード判別システム (RLI) は世界の主要40言語および29種類の文字コードに対応しており、本プロジェクトのニーズにかなう製品であることから採用となりました。

構成図：



導入効果：

RLI の導入により、広範囲にわたる Web データ収集・解析をより正確におこなうことが可能となりました。

早稲田大学工学部 山名早人助教授のコメント：

「本プロジェクトで扱うデータの量は膨大となり、またいろいろな国からの情報を収集し解析するものですから、データの言語および文字コードの判別をおこなうシステムは必要不可欠でした。Web ページの記述言語を判別してくれるこの Rosette 言語・文字コード判別システムは主要なサーチエンジン、さらには、Web データを扱う世界の多数の製品で使用されているデファクト的なものであり、本プロジェクトで採用することにしました。ベイシステクノロジー様とは、本言語判定システムの性能をさらに上げることを目標に共同して研究を進めています。」