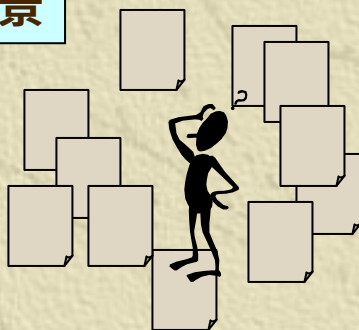


一般化文書頻度計数ライブラリ およびシステムの開発

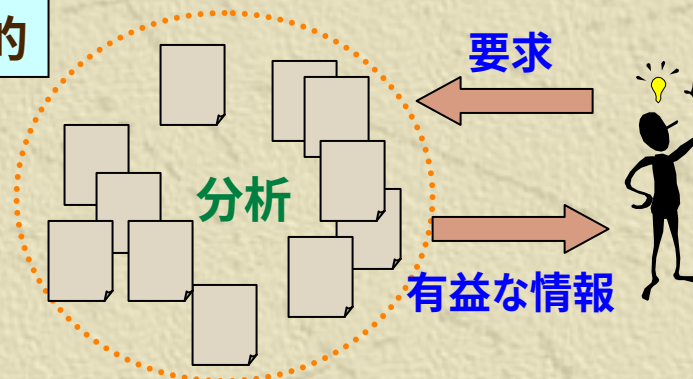
開発者: 寺尾健一郎

背景



大量の文書があっても
必要な情報が発見できない

目的



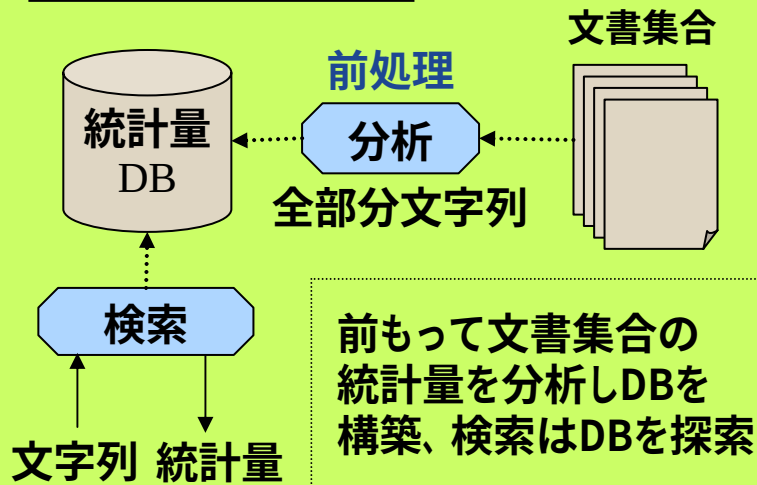
大量の文書を分析することで、
ユーザの望む情報を提供

大量の文書集合から統計量を高速に分析し検索するシステムを開発

システムの特徴

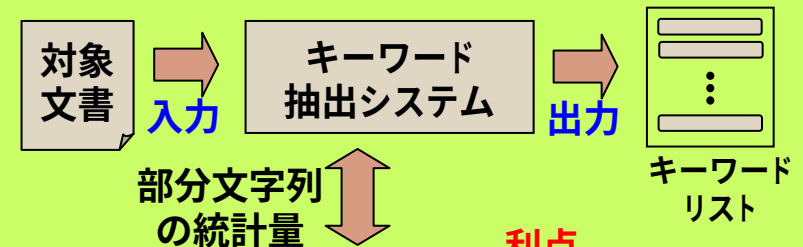
- 10～100GBという巨大な文書集合に対応
- あらゆる文字列の統計量を検索可能
 - ・ 文書集合の全部分文字列を分析。特定単語だけでなく複合語から任意文字列まで
- さまざまな統計量が取得可能
 - ・ 文字列の出現数、出現文書数に加えて文字列がk回以上出現する文書の数も計算
 - ・ 他に、文字列の全ての出現位置、出現文書も得られる

システムの概要



システムの応用事例

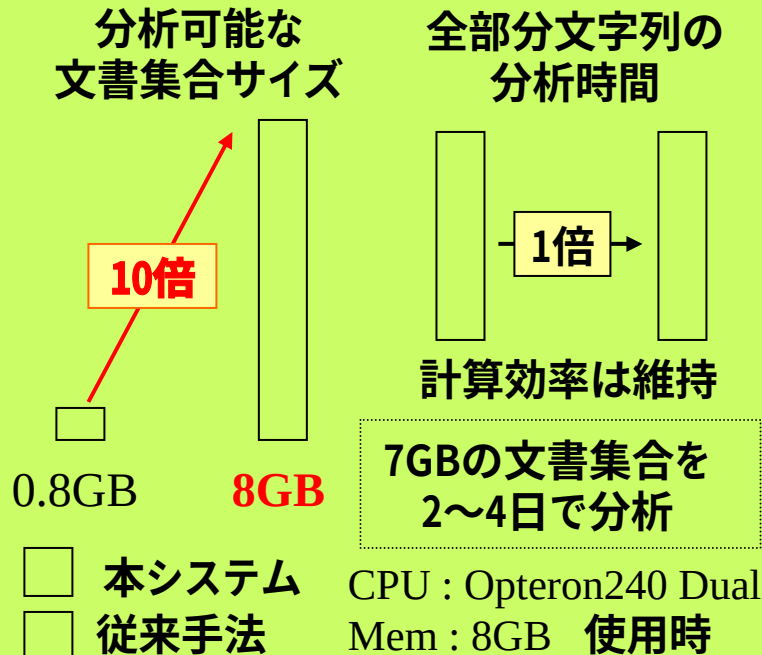
- ・統計情報のみによるキーワード自動抽出システム



利点

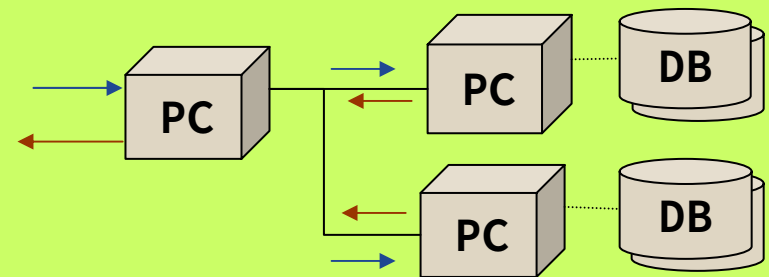
- ・辞書の保守不要
- ・未知語に対応

統計量分析の性能



統計量の検索

- ・複数PCの複数ディスクにDB格納
- ・PCクラスタ上で分散検索し結果を併合
- ・検索は並列に実行



- ・全体で10GB~100GBの文書集合から統計量取得可能