宇宙物理学文献データベースADSからの情報抽出

トラン ヴァン ディエン, 天笠 俊之, 北川 博之（筑波大学）

協力：海老沢研氏，中平聡志氏（JAXA）
大石雅寿氏，市川伸一氏（国立天文台）

背景：学術文献データベースとその活用

学術文献データベース

- 学術論文，書籍等の文献情報およびテキスト情報を含むデータベース．
- 代表的な文献データベース．

dblp.uni-trier.de Computer Science Bibliography　CiNii 日本の論文をさがす Articles　PubMed

ADS (The SAO/NASA Astrophysics Data System)

- 天体物理学分野の文献データベース．
- 1995年から運用．
- 物理学，天文学，天体物理学の990万件以上の文献情報を含む．

学術文献データベースの活用

- 学術文献データベースから，有益な情報を抽出することが可能．

リンク構造解析

- 文献（研究者）の重要度のランキング：PageRank, ObjectRank等

クラスタリング

- 文献（研究者）の分類：最大マージンクラスタリング [Nguyenら]

自然言語処理

- 研究動向の抽出：自然言語パターン解析[難波ら]

複数の分野を含む場合，単一のランキングでは不十分
➔ランキングとクラスタリングの併用

研究の目的：ADSからの情報抽出

- クローラの設計，実装
- 文献データベースからの情報抽出（RankClusの適用）

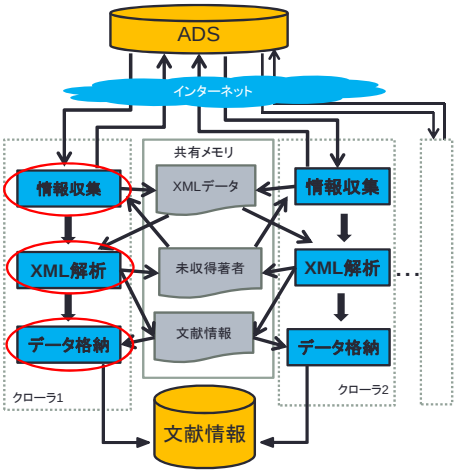
ADSクローラの設計と実装

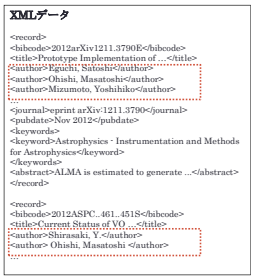
ADSのデータは一括ダウンロード不可
➔クローラによるデータ取得が必要

クローリングの方針

- Web APIの利用．
  - 著者を指定しXML形式でデータを取得可能
- 少数の著者をシードとして，再帰的に共著者を取得．

クローラの概要





取得済みデータ（クローリング継続中）

- 著者数：167,704
- 文献数：2,665,166

文献ID	タイトル	キーワード	概要	リンク
2012ASPC..461..451S	Current Status of ...	Null	In these years, standards ...	http://....; http://...;
2012ASPC..461..375K	VOCrawler...	Null	We report on the...	http://....; http://...;
2012PASA...29..229H	The Completeness and ...	data analysis, radio ...	The process of determining...	http://....; http://...;

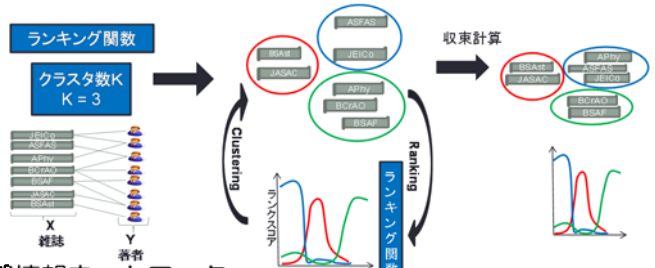
文献ID	研究者ID
2012ASPC..461..451S	1190
2012ASPC..461..451S	1191
...	...
2012PASA...29..229H	1250

研究者ID	研究者の名前
1190	Shirasaki, Y
1191	Komiya, Y
...	...
1250	Huynh, M. T

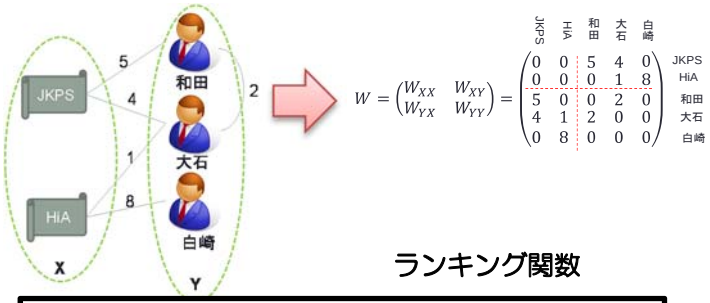
RankClusによる情報抽出

RankClusの概要

- 二部情報ネットワークが対称．
- クラスタリングとランキングを交互に実行．



二部情報ネットワーク



ランキング関数

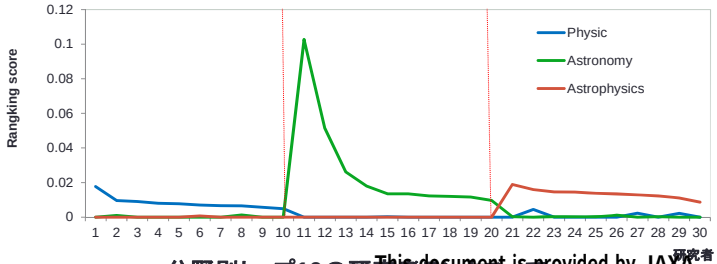
Authority rank

$$\vec{r}_Y(j) = \alpha \sum_{i=1}^{m} W_{YX}(j,i)\vec{r}_X(i) + (1-\alpha) \sum_{j'=1}^{n} W_{YY}(j,j')\vec{r}_Y(j')$$

$$\vec{r}_X(i) = \sum_{j=1}^{n} W_{XY}(i,j)\vec{r}_Y(j)$$

結果

- 3分野34雑誌をランダムに抽出（著者数6,261）



分野別トップ10の研究者ランクスコア