# 探査衛星プロジェクトの評価手法に関する一考察

## -トピックモデルによる論文の要旨分析-

2020年2月14日

2019年度 宇宙科学情報解析シンポジウム

水上 祐治*（日本大学）

高宗 大起（日本大学 学生）・大畠 昭子（JAXA）

中野 純司（中央大学、統計数理研究所）

# はじめに

本研究では、論文の要旨にテキストマイニング分析とトピックモデル分析を施して、一連の研究の変遷を明らかにすることを目指す。

分析では、宇宙開発プロジェクトのX線探査衛星SUZAKUの論文群を題材にしてテキストマイニング分析とトピックモデル分析を行う。

# 目的と研究方法

分析は２段階で行う。

①研究の全体像を示すためのSUZAKU研究の通期をまとめた分析を施す。

②前期（2007年、2008年）、後期（2017年、2018年）の論文種別の詳細把握を行うため、このグループにトピックモデル分析を施す。

# 目的と研究方法

分析では、統計ソフトRを用いて論文要旨にトピックモデル分析を施した。英文のテキストマイング分析には、tmパッケージとNLPパッケージを用いた。そして、トピックモデル分析には、topicmodels パッケージの LDA(Latent Dirichlet Allocation)におけるVEM(Variational Expectation-Maximization)モデルを用いた。

# 分析データ

分析対象は、Ｘ線天文衛星「SUZAKU」である。この衛星は、遠距離にある天体のＸ線観測、宇宙の高温プラズマのＸ線分光観測等を目的としている衛星であり、2005年にM-Vロケット6号機にて内之浦宇宙空間観測所から打ち上げられ、10年間の運用後、2015年に観測を完了している。

# 分析データ

論文情報は、Clarivate Analytics社の書誌データベースWeb of Science core collectionを用いて収集した。表1に論文の検索条件を示す。

表1 Web of Science core collectionでの検索条件

| |
|---|
| WoS 検索条件：(TS=(SUZAKU) OR TS=("ASTRO-E II") OR TS=("ASTRO-E 2") OR TS=("ASTROE II") OR TS=("ASTROE 2") OR TS=("ASTROEII") OR TS=("ASTROE2") OR TS=("ASTRO-EII") OR TS=("ASTRO-E2")) AND 言語: (English) AND ドキュメントタイプ: (Article OR Review) |
| 索引=SCI-EXPANDED, SSCI, A&HCI, ESCI タイムスパン=全範囲 |

# 分析①（全体像把握）

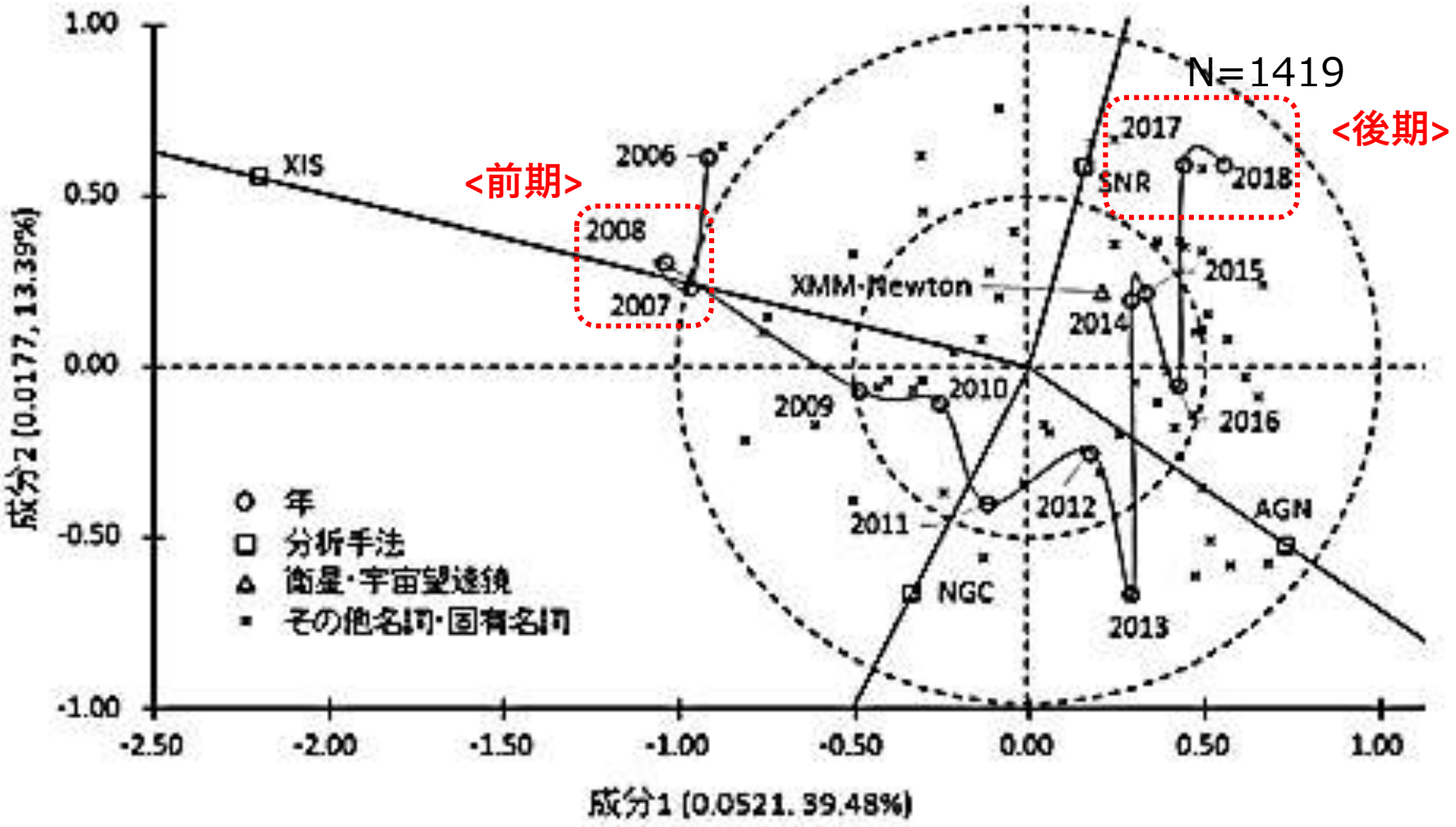



図1 SUZAKUの関連研究のトレンド（コレスポンディング分析）

# 分析②（通期、前期）

まず、テキストマイニング分析にて単語抽出と品詞分類を行った。表に品詞別の単語出現数を示す。次のトピックモデル分析では、名詞(NN)と固有名詞(NNP)のみを用いる。

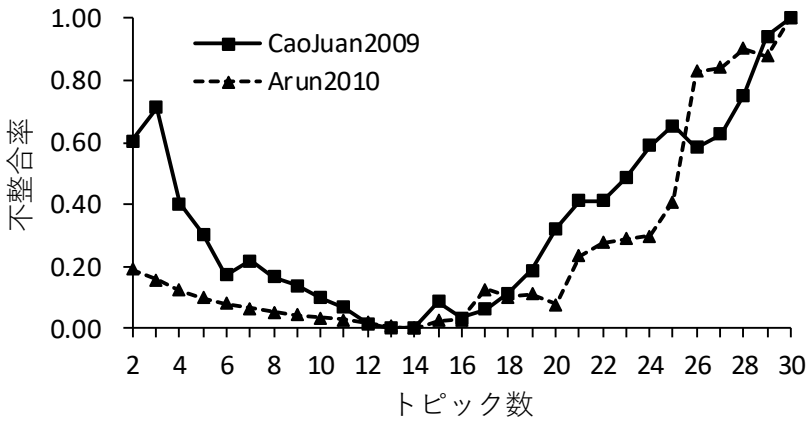
表　品詞別出現数

| | , | . | CC | CD | DT | EX | FW | IN | JJ | JJR | JJS | LS | MD | NN | NNP | NNPS | NNS | PDT | POS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 通期 | 204 | 95 | 526 | 1456 | 580 | 2 | 462 | 5451 | 33587 | 1055 | 384 | 1 | 1133 | 69692 | 505 | 10 | 19011 | 1 | 20 |
| 前期 | 17 | 13 | 66 | 150 | 77 | | 64 | 675 | 3607 | 127 | 50 | | 128 | 7831 | 100 | 1 | 2022 | – | 3 |
| 後期 | 19 | 9 | 56 | 177 | 57 | | 63 | 641 | 4228 | 129 | 52 | | 118 | 8758 | 55 | 2 | 2355 | 1 | 2 |
| | PRP | PRP$ | RB | RBR | RBS | RP | SYM | TO | UH | VB | VBD | VBG | VBN | VBP | VBZ | WDT | WP | WP$ | WRB |
| 通期 | 375 | 10 | 8535 | 229 | 6 | 67 | 111 | 26 | 5 | 2064 | 6667 | 4660 | 7601 | 4125 | 3633 | 50 | 15 | 43 | 28 |
| 前期 | 48 | 1 | 987 | 21 | 2 | 8 | 17 | 4 | – | 217 | 768 | 415 | 876 | 401 | 435 | 7 | – | 5 | 3 |
| 後期 | 45 | | 1057 | 33 | 1 | 6 | 20 | 4 | 3 | 249 | 799 | 598 | 940 | 531 | 437 | 12 | 1 | 2 | 5 |

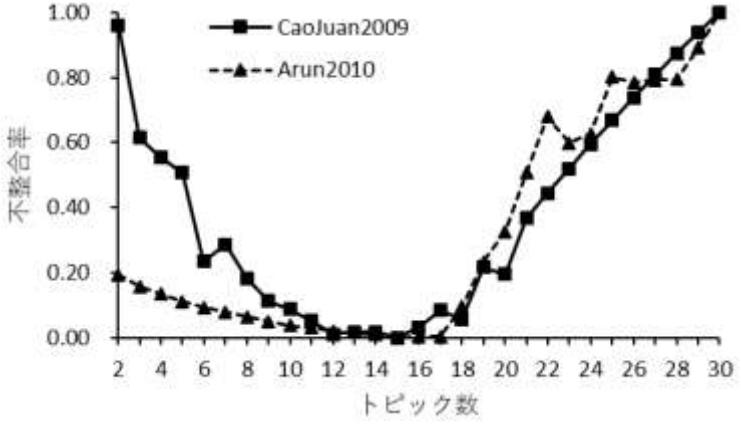
論文数：All：1419編，2007-2008：177編，2017-2018：166編

# 分析②（通期、前期、後期）

**トピック数把握**：CaoJuan2009[8]とArun2010[9]の指標を用いた分析結果を示す。




(a) 前期
(14トピックを選択、N=177)




(b) 後期
(15トピックを選択、N=166)

両指標の不整合率が共に低いトピック数を選択した。

# 分析②（前期 分析結果）

表4 前期のトピックモデル分析の結果　N=177

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | grb | hess | disk | spectrometer | cluster | cyclotron | plasma |
| 2 | continuum | flux | reflection | ecs | abundance | resonance | point |
| 3 | position | tev | iron | detector | ngc | decay | diffuse |
| 4 | pulse | gammaray | component | resolution | metallicity | grb | ionization |
| 5 | mission | telescope | hole | instrument | metal | satellite | field |
| 6 | lag | src | variability | laboratory | galaxy | detector | gammaray |
| 7 | component | coma | powerlaw | cap | telescope | burst | origin |
| 8 | satellite | field | continuum | performance | kpc | phase | temperature |
| 9 | blackbody | wind | accretion | area | icm | pulsar | core |
| 10 | phase | burst | flux | satellite | density | timing | iron |
| | **8** | **9** | **10** | **11** | **12** | **13** | **14** |
| 1 | charge | powerlaw | loop | temperature | cluster | component | plasma |
| 2 | structure | plasma | xis | intensity | iron | xis | xis |
| 3 | method | component | background | flux | velocity | loop | center |
| 4 | transfer | iron | resolution | center | power | nxb | absorption |
| 5 | xis | center | cygnus | plasma | law | ejecta | structure |
| 6 | paper | state | ionization | powerlaw | supernova | limit | photon |
| 7 | device | variability | structure | abundance | temperature | cygnus | spectrometer |
| 8 | chargeinjection | disk | plasma | component | light | chandra | supernova |
| 9 | injection | sgr | satellite | tail | gas | abundance | temperature |
| 10 | excess | cloud | detection | background | center | flux | position |

# 分析②（上位単語の数に関する試行）

トピックの意味を評価者が把握するためには、
固有名詞のように意味が狭い範囲で断定できる
語句が多いことが求められると考えられる。

$$x = ceilin\left(\frac{n_{NN}}{n_{NNP}}a\right) + b$$

$$x = ceilin\left(\frac{8758}{55}0.1\right) + 2 = 18$$

$n_{NN}$ ： 名詞の種類数

$n_{NNP}$ ： 固有名詞の種類数

$a$ ： 定数 (0.1)

$b$ ： 切片 (2)

# 分析②（後期 分析結果）

表6 後期のトピックモデル分析の結果　N=166

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | nustar | state | plasma | gas | accretion | luminosity | outburst | shock |
| 2 | disk | gas | temperature | cluster | mass | flux | iron | cluster |
| 3 | jet | cluster | supernova | mass | component | continuum | pulse | temperature |
| 4 | accretion | accretion | component | galaxy | wind | variability | plasma | hole |
| 5 | neutron | mass | mass | density | spin | component | luminosity | galaxy |
| 6 | component | density | snr | chandra | rate | gammaray | pulsar | number |
| 7 | field | temperature | ionization | enrichment | disk | wind | snr | mach |
| 8 | star | neutron | origin | iron | pulse | neutronization | intensity | acceleration |
| 9 | temperature | luminosity | cloud | chemical | photon | photon | state | jump |
| 10 | blackbody | star | nustar | distribution | discovery | metallicity | blackbody | presence |
| 11 | powerlaw | brightness | process | matter | plasma | reflection | feature | detection |
| 12 | flux | log | shock | icm | nustar | evolution | phase | polarization |
| 13 | scenario | distribution | electron | supernova | presence | progenitor | flux | structure |
| 14 | luminosity | halo | cie | value | jet | xi | absorption | comptonization |
| 15 | plasma | core | gas | evolution | evolution | gev | temperature | accretion |
| 16 | quasar | lbol | galaxy | abundance | absorption | outflow | gammaray | disk |
| 17 | xmmnewton | edge | structure | nustar | remnant | ionization | ax | component |
| 18 | chandra | pressure | rim | group | origin | nustar | presence | front |
| 19 | formation | nustar | gev | formation | context | absorption | component | telescope |
| 20 | state | presence | halo | intracluster | reflection | ratio | supernova | mechanism |

# 分析②（後期 分析結果つづき）

表7 後期のトピックモデル分析の結果　N=166

| | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|
| 1 | variability | reflection | loop | temperature | background | reflection | nustar |
| 2 | torus | disk | outflow | shock | velocity | iron | temperature |
| 3 | chandra | xmmnewton | system | cluster | gas | accretion | hole |
| 4 | ionization | iron | plasma | distance | core | component | mass |
| 5 | density | component | star | gammaray | cluster | continuum | depth |
| 6 | absorption | nustar | disk | plasma | km | hole | seyfert |
| 7 | archival | continuum | rate | direction | spectrometer | variability | catalogue |
| 8 | compton | instrument | wind | mpc | measurement | luminosity | xmmnewton |
| 9 | telescope | accretion | density | matter | find | outflow | literature |
| 10 | galaxy | density | gas | density | blackbody | absorption | ratio |
| 11 | abundance | hole | dotout | position | component | seyfert | photon |
| 12 | correlation | value | arc | field | agn | galaxy | level |
| 13 | gravity | powerlaw | ridge | pulsar | orbit | density | comptonization |
| 14 | lx | comparison | interpretation | nebula | hitomi | bulk | rate |
| 15 | knot | fraction | galaxy | extent | galaxy | excess | significance |
| 16 | uv | inclination | continuum | presence | supernova | abundance | corona |
| 17 | behavior | response | powerlaw | structure | continuum | comptonization | correlation |
| 18 | angle | ser | accretion | galaxy | perseus | structure | work |
| 19 | hole | ring | photoionization | front | ism | velocity | context |
| 20 | reflection | constraint | order | interpretation | xis | powerlaw | hand |

# まとめ

本研究では、論文の要旨にテキストマイニング分析とトピックモデル分析を施して、一連の研究の変遷を明らかにすることを目指した。

分析では、宇宙開発プロジェクトのX線探査衛星SUZAKUの論文群を題材にしてトピックの解釈をおこなった。

# まとめ

本研究では、全体像を示すためにテキストマイニング分析を行い、各年のさらに詳細な区分を把握するために、トピックモデル分析を行った。また、分析後のトピックの解釈を容易にすることを目的として、トピックの特徴を示す単語を下位まで調査する等の改良手法を示した。

# まとめ

分析では、一般に入手可能な情報を用いて、X線探査衛星SUZAKU論文群の前期、後期での研究のトレンドを示すことができた。

今後は、分析で抽出したトピックと専門家の理解の差を把握して、それを小さくするための研究を行う予定である。

# 参考文献


[1] Yuji Mizukami, Keisuke Honda and Junji Nakano, Study on Research Trends on the Internet of Things Using Network Analysis, IJAMS, Vol.10, No.1, pp.27-35 ,2018

[2] 根岸正光, 山崎茂明, 「研究評価―研究者・研究機関・大学におけるガイドライン」, 丸善、2001

[3] 伊藤慎太朗、水上祐治、「トピックモデル分析を用いたビックデータ研究の動向分析」、日本経営システム学会　第63回　全国大会研究発表大会　講演論文集、pp. 228-229, 2019

[4] Select number of topics for LDA model <https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html> (2019/10/1)

[5] 松河秀哉、大山牧子、根岸千悠、新居佳子、岩﨑千晶、堀田博史、「トピックモデルを用いた授業評価アンケートの自由記述の分析」、日本教育工学会論文誌 41(3)，233-244，2017

[6] 黒宮寛之、日高一郎、山本義春、「トピックモデルによる研究型アクティブラーニングの分析」、日本教育工学会論文誌 42(4)，323-330，2019

[7] 中村匡佑、水上祐治、大畠昭子、「テキストマイニング分析による探査衛星の運用成果に関する一考察　–X線探査衛星すざくの関連論文を題材として–」、第62回 日本経営システム学会 全国研究発表大会講演論文集、pp.236-237

[8] Cao Juan, Xia Tian, Li Jintao, Zhang Yongdong, and Tang Sheng. 2009. A density-based method for adaptive lda model selection. Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008 72, 7–9: 1775–1781.

[9] Rajkumar Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In Advances in knowledge discovery and data mining, Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran and Vikram Pudi (eds.). Springer Berlin Heidelberg, 391–402.