報道発表

- **数値・固有名詞情報の抽出可視化を商用で業界初の実現**
  ～テキスト情報からマーケティング情報や社会動向を簡易に摘出・図式化が可能に！ ～

- **平成20年9月30日**

独立行政法人情報通信研究機構(以下「NICT」という。理事長：宮原 秀夫)の言語基盤グループの村田主任研究員他は、株式会社数理システム（代表取締役社長：山下 浩）と共同で、Web上のニュースサイトなどの膨大なテキスト情報から、任意の分野における固有名詞（商品・サービス名称など）とそれに関連する数値情報（価格、数量、時刻など）を自動抽出し、それらをコンピュータ上で可視化することができるテキストマイニングシステムを開発しました。これにより、Webのテキスト情報をグラフ化して分析することができ、マーケティングの分析などに役立ちます。今後、NICTは、同開発システムを株式会社数理システムに技術移転し、同社が開発・販売しているテキストマイニングシステムに実装し、実用化する予定です。

【背景】

テキストマイニングシステムとは、これまで人が読んで理解してきたテキスト情報を、大量かつ迅速に把握して知識を採掘（Mining）するコンピュータシステムのことで、テキスト情報の頻度分析、時系列分析などを行い、文書内容の特徴を把握するものです。単に、テキスト情報にある単語を単体で扱うのではなく、形態素構文解析から単語の品詞や係り受けの関係の情報が抽出できます。単語抽出の性能は、辞書機能に依存しています。こうしたテキストマイニングシステムには、Web上のニュースや文章などのテキスト情報から、ある商品やサービスに関係した価格、支持率、時刻、数量などの数値情報の関連抽出のニーズがあります。しかし、ニュースにおける商品・サービス名称といった時事に関する固有名詞は、リアルタイムでの辞書作成が困難であり、また、固有名詞と数値情報の関連性を抽出することは不十分でした。

【今回の成果】

今回開発したシステムは、NICTの自然言語処理技術*1を応用し、単位付き数値情報と固有名詞を、その単位及び種別ごとにまとめることで、関連性を抽出、グラフ化できるという画期的なシステムとなっています。商用システムでは業界初であり、当該技術は特許出願中*2です。

固有名詞と数値情報については、文中に存在するこれら単語間の位置に基づく距離から関連付けを行い、システム搭載の固有名詞コーパス*3から作成したモデルに基づく類似表現を抽出することができます。また、関心のある分野については、コーパス作製技術*4を利用して、オリジナルのコーパスを用意することで、オーダーメイドによる高精度のテキストマイニングが可能となりました。

開発したシステムは、数理システム社が開発販売している商用のテキストマイニングシステム（Text Mining Studio）のオプション機能として動作します。例えば、石油製品の価格変動に関する話題など関心のあるマーケティング情報などをweb上の情報を利用して分析し、表やグラフで可視化できます。膨大なニュース情報をリアルタイムで情報分析をすることにより、意思決定の支援ツールとして期待されます。

【今後の展望】

今後は、抽出したデータのグラフから元のデータを直接参照できる機能を搭載する予定です。また、今回開発したテキストマイニングシステムは、NICTの技術移転として、株式会社数理システムがシステムの販売やカスタマイズ、オーダーメイドのコーパス作製等のサービス提供を行います。

なお、幕張メッセで開催される「CEATEC JAPAN 2008」において10月2日（木）の「NICT技術シーズ説明会」で本技術の紹介を行う予定です。

＜ 本件に関する 問い合わせ先 ＞
知識創成コミュニケーション研究センター
言語基盤グループ 村田 真樹
Tel:0774-95-1332
Fax:0774-95-1308

＜ 広報 問い合わせ先 ＞
総合企画部 広報室
報道担当
Tel:042-327-6923
Fax:042-327-7587

＜用語 解説＞

1 **自然言語処理技術**

日常で人間が使う言語(例:日本語)を自然言語といいます。その自然言語をコンピュータで処理することを自然言語処理といいます。自然言語処理の例として、仮名漢字変換や機械翻訳がありますが、テキストマイニングも自然言語処理の一種です。

*2 **特許出願中**

● 数値固有名詞情報の抽出とそのグラフ化の特許(3件)
- 特願2007-130218、 情報抽出装置、情報抽出方法及び情報抽出プログラム
- 特願2006-191076、 情報抽出装置、情報抽出方法及び情報抽出プログラム
- 特願2006-016052、 情報抽出・表示装置、情報抽出・表示方法及び情報抽出・表示プログラム

● コーパス作製に関する特許(1件)
- 特許第3396734号、 コーパス誤りの検出・修正システム，コーパス誤りの検出・修正処理方法及びそのプログラム記録媒体

*3 **コーパス**

テキスト情報を収集したデータのことをコーパスといいます。テキスト情報に、注釈を付けたデータをタグ付きコーパスといいます。例えば、固有名詞コーパスでは、テキストにどの箇所が人名か地名か組織名かという注釈がふられたコーパスです。

*4 **コーパス作製技術**

タグ付きコーパスを人手で効率よく作成するための技術です。コーパスの作製には、コーパス修正技術*5が役立ちます。

*5 **コーパス修正技術**

タグつきコーパスを人手で効率よく作成するための技術です。コーパスは人手で作成するために誤りが生じます。その誤りを修正するための技術です。コーパスの修正には、最大エントロピー法*6と呼ばれる機械学習の方法が役立ちます。

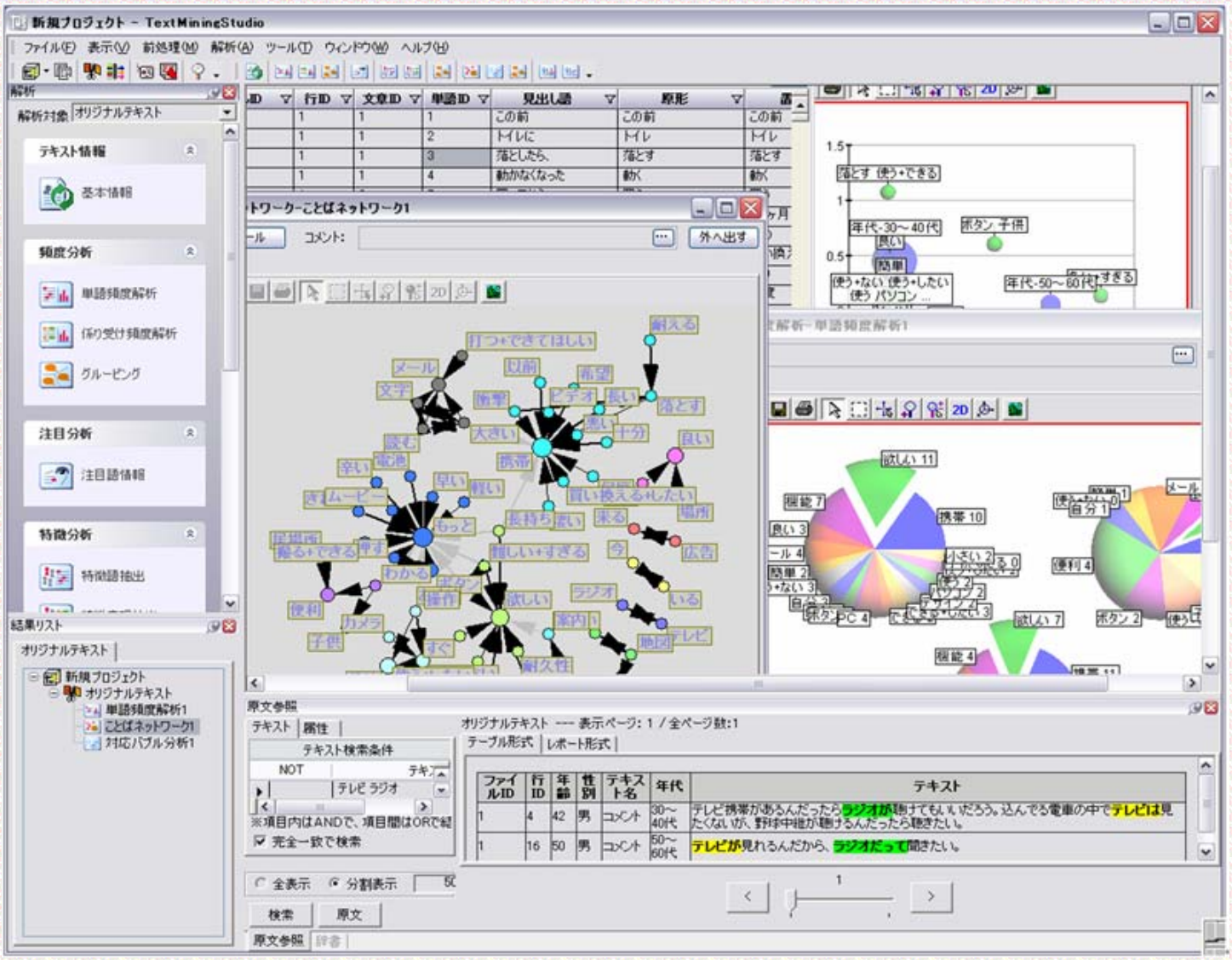
*6 **最大エントロピー法**

最大エントロピー法は、学習データでの素性(推定に用いられる情報の細かい単位のこと)の出現の期待値と未知データでのそれが等しいという条件で確率分布のエントロピーが最大の場合の確率分布を求め、それに基づき素性の各出現パターンに対して各分類になる確率を求め、分類先を推定するものです。それぞれの分類先が正しい確率が求められるため、コーパス中の各データが誤っている確率を求めることができ、コーパス中の誤りを分析するのに便利な手法です。

補足資料

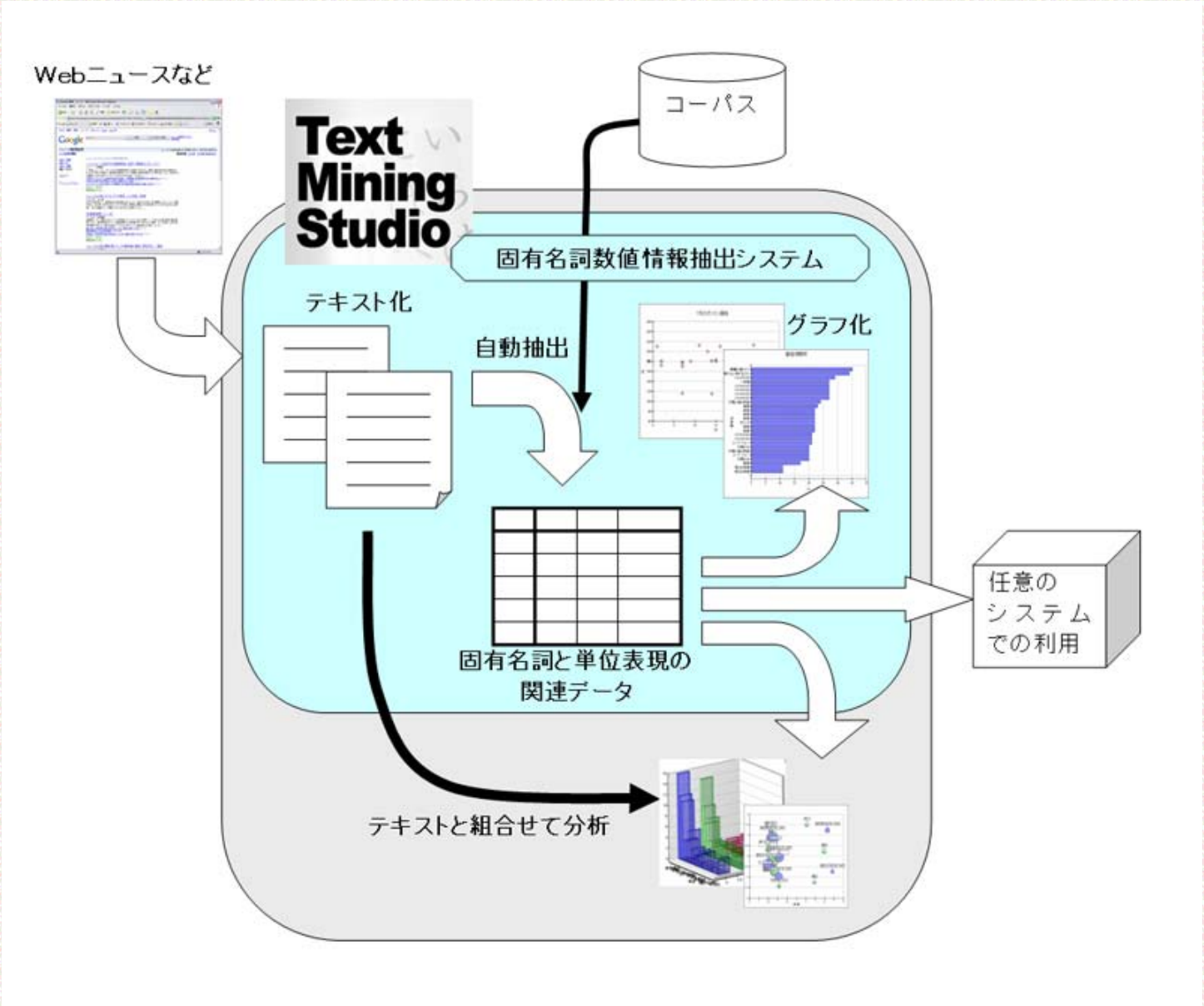
＜テキストマイニングシステム図＞

これまで人が読んで理解してきたテキストデータを、大量かつ迅速に把握して知識を採掘（Mining）するコンピュータシステムのことで、テキスト情報の頻度分析、時系列分析などを行い、文書内容の特徴を把握するものです。下の図は、テキストマイニングシステム(Text Mining Studio)の画面の例です。



＜システム概念図＞

今回構築したシステムの概念図を下に示します。従来からある株式会社数理システムのテキストマイニングシステム(Text Mining Studio) に、合体可能な形でシステムを構築しました。Webニュースなどが表示しているhtmlファイルなどから簡便にテキストを抽出することが可能になっています。また、そのテキストから単位付きの数値情報と固有名詞を、その単位ごと及び種別ごとにまとめあげた関連データを作成できます。その関連データをグラフ化することも可能です。その他、本システムにより取り出したテキストや固有名詞と単位表現の関連データをテキストマイニングシステム(Text Mining Studio)でより詳しく分析することもできますし、任意のシステムで利用することもできます。テキストからの単位付き数値情報と固有名詞の取り出しは、システム搭載の固有名詞コーパスから作成したモデルに基づき行います。コーパスには、取り出したい正しい固有名詞や数値情報が記載されており、それを計算機がモデルを学習することで、数値情報と固有名詞を取り出すことができるようになります。



＜実行例＞

例えば、下記のようなデータを用いて抽出、分析することができます。

ガソリン価格に関する実際のWebニュース記事の一部

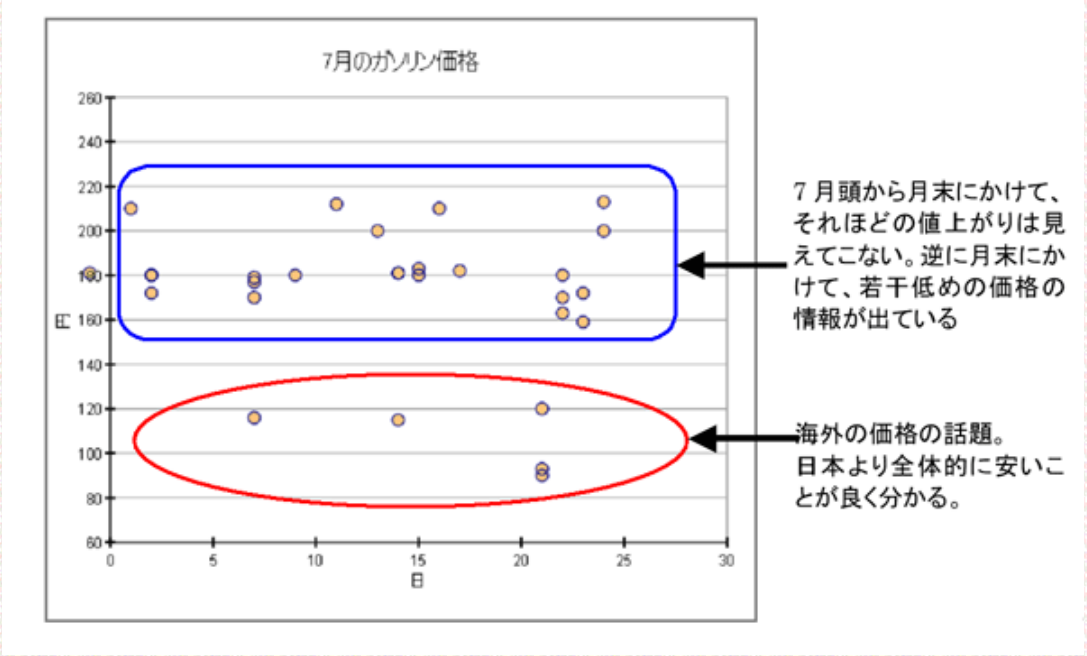
（グーグルニュースhttp://news.google.co.jp/より抜粋。ニュース検索ワード「ガソリン　リットル 円」）

8月は180円台後半も＝ガソリン卸値、6円値上げ示唆－石連会長。時事通信 - 2008年7月24日。石油連盟の天坊昭彦会長（出光興産社長）は23日の定例記者会見で、石油元売り大手の8月1日からのガソリン卸値に関し、原油価格が現状のままの場合「（1リットル当たり）6円前後の値上がりになる」との見通しを示した。卸値の引き上げが実現すれば、レギュラー ・・・

都市値下げ合戦 町村まだ高値／ガソリン迷走 業界悲鳴。沖縄タイムス - 2008年7月24日。止まらぬ原油高騰と競争激化で、ガソリン価格が迷走している。本島中南部の都市部では、今月中旬から一気に相場が下がり、那覇近郊の店では二十三日、一リットル百五十九円を付けた。一方、離島や町村部は高値傾向が続き、八重山の離島では二百三円も ・・・

2週連続の値下がり レギュラーの平均価格。47NEWS - 2008年7月24日。石油情報センターが24日発表した石油製品市況の週間動向調査（22日現在）によると、レギュラーガソリンの全国平均小売価格は、前週に比べ1リットル当たり40銭安の180円90銭と、2週連続で値下がりした。 小売価格の上昇が続いて消費者が買い控える動きが出 ・・・

上記の文章から数値を抽出して可視化して表示した例



ニュース記事から抽出したガソリン価格データのグラフ

また、テレビの番組名のコーパスを別途作成することにより、各ニュースで論じられている番組視聴率データを取得することも可能となります。

視聴率に関する実際のWebニュース記事の一部

（グーグルニュースhttp://news.google.co.jp/より抜粋。ニュース検索ワード「ドラマ　視聴率」）

朝日新聞 - 2008年7月22日。NHK大河ドラマ「篤姫」が人気だ。薩摩藩から徳川将軍家へ嫁ぎ波乱の生涯を送った幕末の女性を宮崎あおいが演じる。20日までの平均視聴率は、23%で大河では過去11年間で最高だった「毛利元就」（97年）の年間平均視聴率と並んだ（関東地区、ビデオリサーチ ・・・

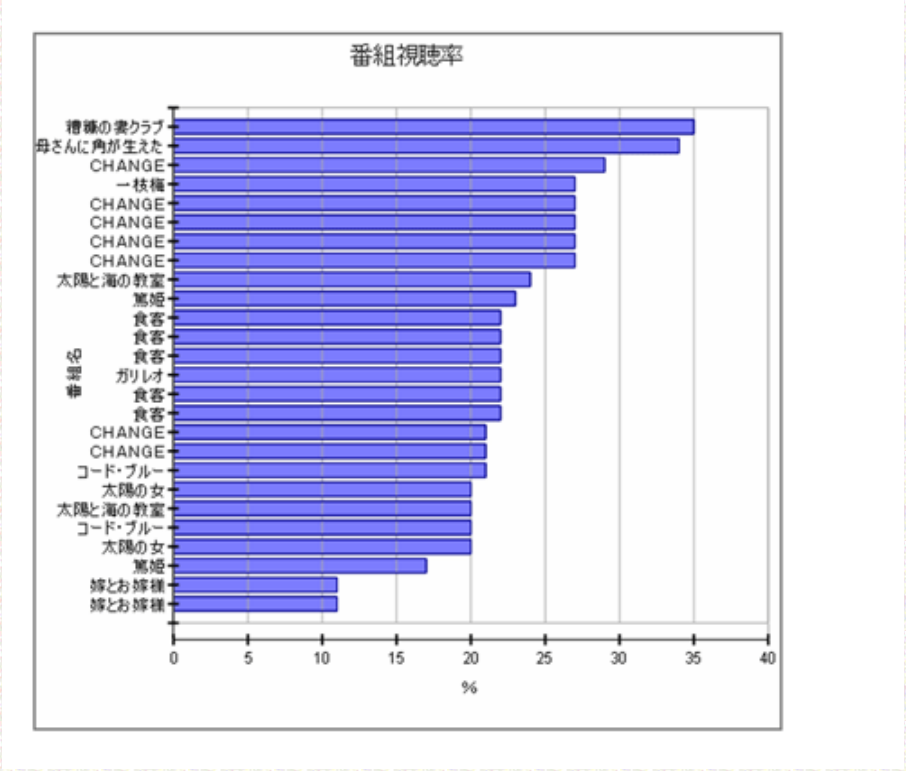
日経トレンディネット - 2008年7月22日　前クールの『ライフ』はほかの7月クールのドラマと同様、9月中旬に終了したものの、間に単発ドラマを挟むなどして、『SP』がスタートしたのはほかの10月クールドラマから大きく遅れること11月

3日。が、そんな半端な開始時期が功を奏したのか、初回視聴率は15%と高視聴 ・・・

【視聴率】『一枝梅』27% ・・・有終の美を飾れるか innolife.net

視聴率 ：『太陽の女』絶好調20% 朝鮮日報

上記の文章から数値を抽出して可視化して表示した例



テレビ番組名コーパスを作成した場合の番組視聴率データのグラフ

なお、コンピュータ画面に表示された特定のグラフからそのグラフの基になったニュース素材を表示させることを予定しています。