Uma análise da similaridade dos Discursos Parlamentares

RESUMO

A análise de dados governamentais é observada em diversas áreas, mas poucas são as análises encontradas em conjuntos de dados textuais do governo. Neste artigo foram analisados 38035 discursos de parlamentares com o objetivo de identificar similaridade entre partidos e entre os próprios parlamentares. A análise foi realizada através de duas abordagens. Além disso, foram utilizadas técnicas de visualização da informação para facilitar o entendimento dos dados extraídos através de técnicas de mineração de texto.

Palavras-Chave

Busca e recuperação da informação, visualização da informação

ABSTRACT

Categories and Subject Descriptors

H.4.0 **[Information System Applications]**: General

General Terms

Management, Human Factors.

Keywords

Information retrieval, visualization information

1. INTRODUÇÃO

O ano de 2016 foi um ano singular na política brasileira, quando ocorreu o quarto impeachment presidencial. Diversos argumentos e análises foram apresentados como justificativas para condução do processo. No entanto, não foram encontrados na área de recuperação da informação, trabalhos que tenham sido desenvolvidos baseados nos textos dos discursos da câmara dos deputados, de forma a conduzir parlamentares em uma ou outra direção.

O ano foi especial também na nítida mudança de postura do maior partido aliado na base governista, que atuou de forma chave para conduzir o impeachment da presidente. Dessa forma podemos levantar algumas questões como: Será que essa mudança de discurso poderia ter sido detectada de forma antecipada por um algoritmo? Assumindo que discursos semelhantes possam ser ligados a parlamentares, poderia um algoritmo traçar semelhanças e diferenças baseado exclusivamente no vocabulário utilizado por cada político?

Dentro desse contexto, seria interessante analisar textos dos discursos da câmara dos deputados com o objetivo de recuperar informações não tão evidentes no conjunto de dados.

Nesse sentido, em conjunto com técnicas de recuperação da informação e mineração de texto, a visualização da informação pode auxiliar o entendimento dos dados facilitando a obtenção do conhecimento através da percepção visual. A InfoVis, como é comumente chamada, foi definida por Dario [2] como um campo que estuda o uso de representações visuais de dados, fazendo uso da capacidade de percepção visual humana, para detectar padrões e tendências. Gomes e Tavares [5] apontaram como vantagens na utilização da InfoVis a imediata percepção da informação, o realce visual das características das informações, a indicação de padrões e relações entre informações, e que as informações incorretas são facilmente evidenciadas.

Além disso, A InfoVis pode acelerar a percepção de informações vindas da grande quantidade de dados abrindo uma grande vantagem na tomada de decisão [7].

Alinhadas a este cenário, há várias iniciativas cujo objetivo é analisar dados de governo. Por outro lado, ainda há muitas questões que necessitam ser investigadas principalmente em relação a análise de dados textuais. O objetivo desse trabalho foi realizar uma análise dos discursos de parlamentares brasileiros e investigar a similaridade. Para isso, foram analisados 38035 discursos de 585 parlamentares.

O artigo está organizado da seguinte maneira: na seção 2, será apresentada a metodologia utilizada para a análise. Na seção 3, será apresentado os experimentos, que foram executados seguindo duas abordagens diferentes. Além disso, na seção 3 será apresentada as técnicas de InfoVis utilizadas para facilitar análise dos dados. Por fim, na seção 4 será apresentada a conclusão.

2. METODOLOGIA

A análise foi executada em cinco etapas, como ilustrado na Figura 1.

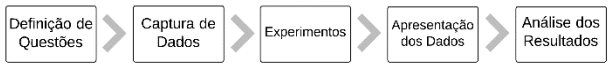

Figura 1. Etapas

Inicialmente foram definidas questões que poderiam gerar resultados relevantes. As questões definidas estão apresentadas na Tabela 1.

Tabela 1 – Questões definidas

Q1	Quais parlamentares possuem discursos semelhantes?
Q2	Quem são os parlamentares com o discurso diferente da maioria de seu partido?
Q3	Os discursos dos parlamentares diferentes da maioria de seu partido se alinham com o de algum outro partido?
Q4	Os discursos se alinham entre partidos diferentes?
Q5	Quais são as palavras que caracterizam e as que descaracterizam partidos e parlamentares?

A captura de dados foi feita através do site da Câmara dos Deputados [3]. Foram utilizadas as atas de reuniões parlamentares, digitadas por taquígrafos. Em uma análise inicial, foi questionado se seria possível responder as questões definidas devido à ausência de informações capazes de caracterizar as dimensões de interesse.

Foi coletado discursos durante sessões da Câmara dos Deputados extraídos pelo webservice no período de 01/01/2015 até

09/06/2016, totalizando 38035 discursos provenientes de 585 parlamentares diferentes e integrantes de 30 partidos possíveis.

3. EXPERIMENTOS

Para realização de busca e recuperação das questões definidas inicialmente, foram aplicadas duas abordagens distintas, de forma a comparar os resultados obtidos por ambas de maneira independente. As abordagens utilizadas foram aprendizado de características e fatoração da matriz principal.

3.1 Aprendizado de Características

Foi utilizado aprendizado de máquina para identificar os termos mais importantes que caracterizam um indivíduo. Para vetorização do texto, foi utilizada a técnica *bag-of-words* ou saco-de-palavras.

Na técnica, para um documento d, o conjunto de pesos determinados por uma função de frequência, que mapeie o número de ocorrências de seus termos t em um real positivo d, pode ser visto como uma síntese quantitativa daquele documento, conhecida como saco-de-palavras [9]. A ordem em que os termos ocorrem no documento não é importante nesta modelagem, o que não impede de ser uma boa representação para comparar documentos com conteúdo semelhantes.

No caso da análise dos discursos, cada vetor, representa um discurso individual de um deputado. O algoritmo de aprendizado utilizado foi a regressão logística multinominal, ou "regressão *softmax*" [13], uma generalização da regressão logística para o problema de classificação entre múltiplas classes. Este modelo foi utilizado devido a sua simplicidade, o que possibilitou o uso de seus coeficientes finais como indicativos da importância relativa de cada atributo. O algoritmo foi treinado para prever a quem pertence o discurso, incentivando o modelo a ter pesos maiores em palavras altamente discriminativas para um determinado parlamentar ou partido.

Os discursos foram pré-processados inicialmente com o objetivo de remover o máximo possível de ruído, como intervenções de outros participantes e observações feitas pela equipe de taquigrafia, já que o formato final proveniente é um RDF não-estruturado.

Um segundo passo de pré-processamento foi a *tokenização*, seguida de *stemmização* das palavras, utilizando o algoritmo de RSLP, acrônimo para Removedor de Sufixo da Língua Portuguesa [10]. O objetivo deste passo é reduzir o léxico que o algoritmo terá de tratar, tentando agrupar palavras diferentes porém com mesmo significado, como verbos conjugados. A palavra mais frequente foi utilizada como representantes destes agrupamentos para as análises.

Outro passo foi a remoção de palavras baseadas na sua frequência relativa por documentos: se uma palavra ocorre em muitos discursos simultaneamente, como pronomes ou artigos, esta palavra pouco irá caracterizar um falante. Por outro lado, se ela ocorre poucas vezes, o resultado será menos interessante pois não haverá comparação relativa entre as outras pessoas. Para este trabalho foram removidas as 100 primeiras palavras mais frequentes em diferentes discursos e utilizadas as 3000 seguintes para análise. Estas frequências podem ser visualizadas na Figura 2. Também foram removidos os nomes dos deputados e dos Estados do Brasil dos discursos analisados, pois estas palavras são pouco interessantes para a análise final e potencialmente muito discriminantes dos parlamentares ou partidos.

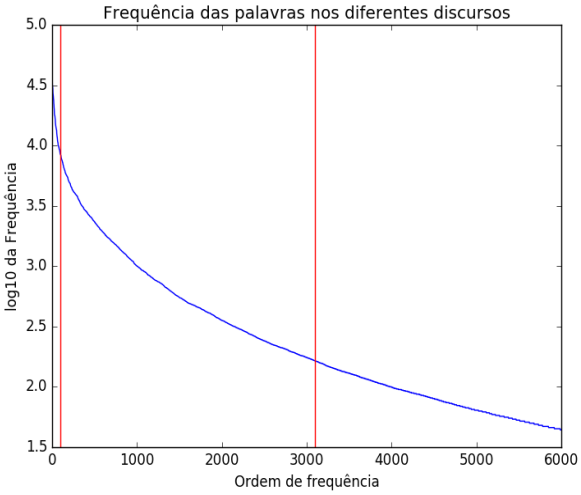

Figura 2. Frequência relativa das palavras

Por fim, para treinar o modelo foi utilizado como entrada do problema de aprendizado supervisionado os vetores do modelo de *bag-of-words* normalizados com a norma euclidiana l_2. Como alvo do treino, foram utilizados os parlamentares que tinham mais do que 50 discursos, totalizando em 203 classes no caso do modelo de políticos, 24 para partidos e uma amostra de tamanho 28670, depois de removidos os discursos que não passaram no crivo. O algoritmo passou pelo processo de validação cruzada para a escolha de seus parâmetros de regularização, sendo executado o algoritmo *k-Folds*, usando a acurácia como critério e três *folds* para cada um dos cinco coeficientes de regularização testados: 10, 1, 0,1 e 0,01.

Para realizar as análises qualitativas, foi gerada uma matriz de similaridade entre os diferentes políticos, utilizando a similaridade dos cossenos com os vetores resultantes do aprendizado. A matriz de similaridade foi utilizada para clusterizar as diferentes classes, utilizando o algoritmo *Affinity Propagation* [4]. Os parâmetros do algoritmo, dumping e preferência, foram selecionados via *silhouette score*, medida de qualidade de clusterização que leva em consideração a coesão entre os elementos de um cluster e a separação deles com os demais clusters [14]. Tais valores podem ser visualizados nos gráficos das Figura 3Figura 4, para partidos e políticos, respectivamente. Com base nestes dados foram escolhidos os valores 0,9 e 0,65 para as constantes de *damping* do algoritmo de *Affinity Propagation* e 0,9 como o valor do parâmetro de preferência em ambos casos, já que para esta medida, quanto mais próximo de 1 melhor.

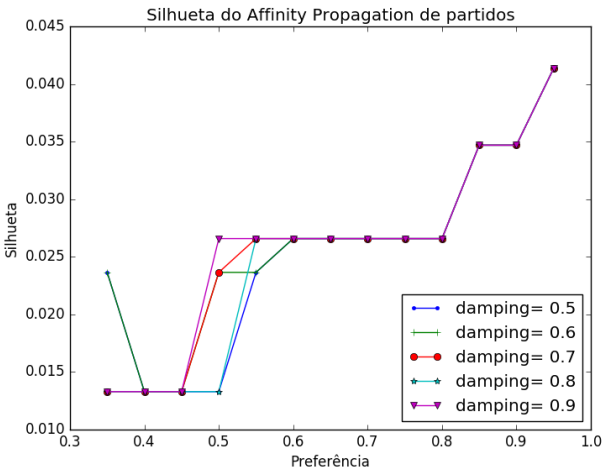

Figura 3. *Silhouette score* para partidos

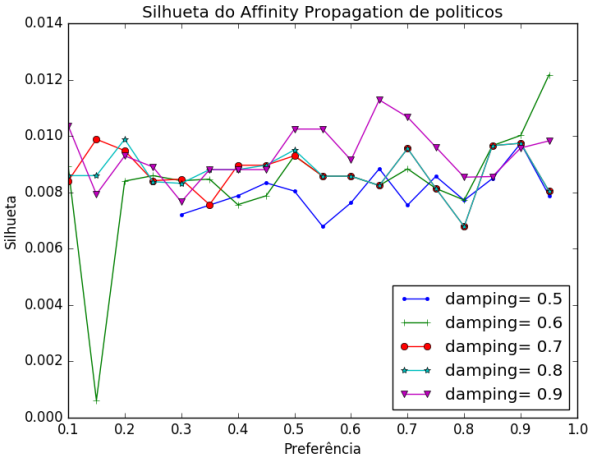

Figura 4. *Silhouette score* para políticos

Para melhor entendimento e análise das informações, os dados foram apresentados em técnicas de visualização de informação.

A técnica nuvem de palavras foi utilizada para evidenciar as palavras caracterizantes e as palavras descaracterizantes, para cada partido, cluster de partidos, parlamentares e cluster de parlamentares. Na Figura 5 é possível observar as nuvens de palavras caracterizantes e descaracterizantes de um cluster de partidos, possibilitando fazer uma comparação.

Figura 5. Nuvem de Palavras

A técnica círculos hierárquicos, foi utilizada para apresentar os clusters. Foram criadas duas visualizações, uma que apresenta os clusters de partidos e uma que apresenta os clusters por parlamentares. Na Figura 6 é possível observar os clusters de partidos.

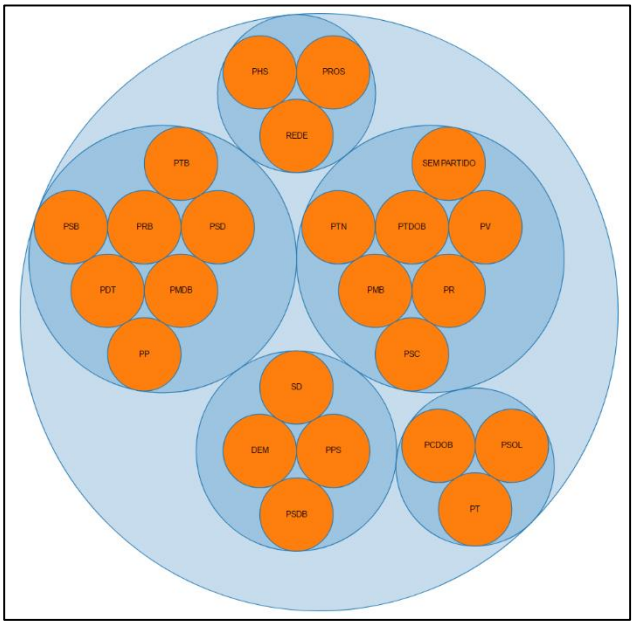

Figura 6. Círculo Hierárquico – Cluster de Partidos

Na Figura 7 é possível observar a matriz de adjacência que apresenta a co-ocorrência da similaridade entre os partidos. Na

caixa de seleção *drop-down* é possível alterar a forma de ordenação das colunas e linhas, por cluster ou por nome do partido. A técnica mapa de calor, também foi utilizada para apresentar a matriz de adjacência que apresenta a co-ocorrência da similaridade entre os parlamentares.

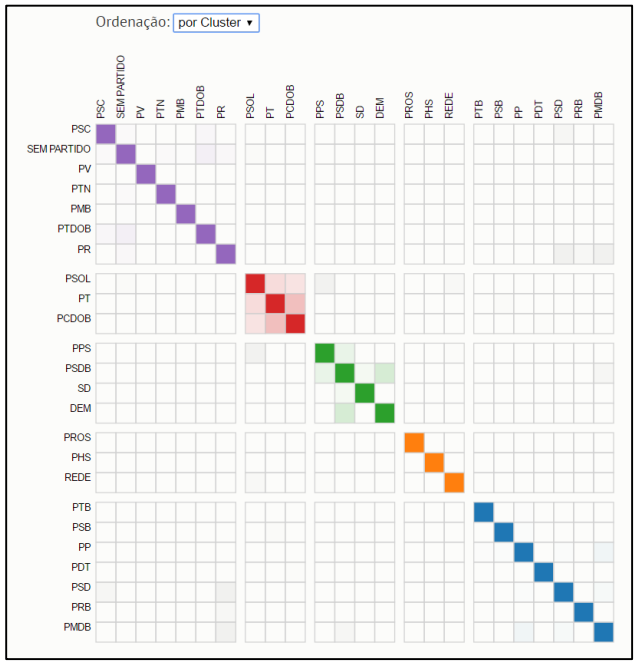

Figura 7. Mapa de calor ordenado por cluster

A matriz de adjacência que apresenta a co-ocorrência da similaridade entre os partidos e políticos também foi apresentada através da técnica diagrama de acordes. Na Figura 8 é possível observar a similaridade dos discursos entre os partidos. É importante destacar que além de apresentar a visão geral dos dados, a técnica é interativa e ao passar o mouse em cima de um dos nós(partidos) é apresentado uma visão mais detalhada do dado como apresentado na Figura 9.

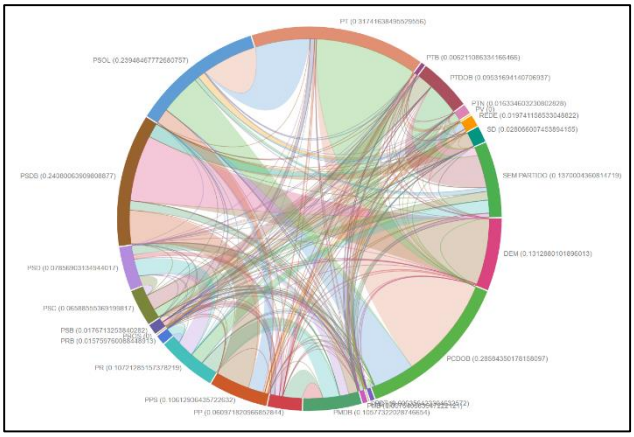

Figura 8. Diagrama de acordes - Similaridade entre partidos

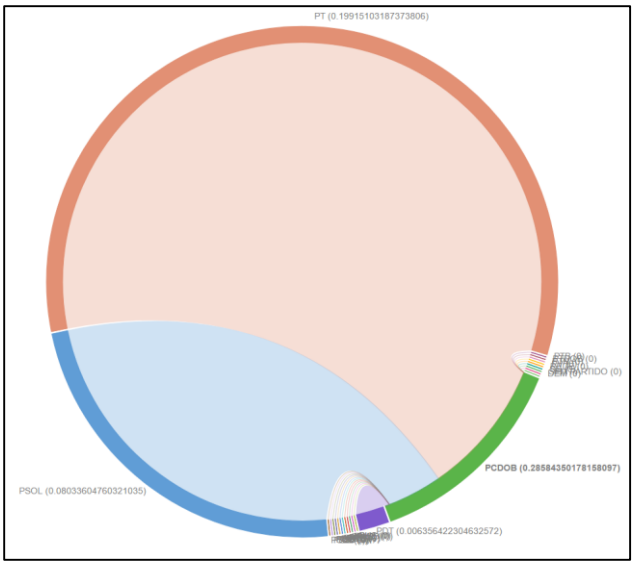

Figura 9. Visão detalhada do diagrama de acordes - Similaridade entre partidos

3.2 Fatoração da Matriz Principal

Utilizando a mesma massa de dados criada na primeira abordagem, formulamos uma abordagem alternativa de forma a confrontar os resultados iniciais. A ideia é verificar se a informação obtida pela aplicação da estratégia TF-IDF sobre os discursos representados em *bag-of-words* é capaz de capturar informações mais significativas na análise da relação parlamentares-parlamentares do que se aplicadas estratégias tradicionais de outras áreas dedicadas à procura de padrões, como na área de Descoberta do Conhecimento (KDD).

Uma estratégia bastante utilizada na área é a fatoração de matrizes, que provê um particionamento da matriz principal em duas ou mais matrizes, de forma a obter vetores de fatores latentes sobre as linhas (parlamentares) e colunas (palavras). Na Figura 10 é possível observar um exemplo utilizado no trabalho de Gower [6]. A partir daí o problema pode ser analisado no espaço de representação composto pela matriz de fatores, abrindo possibilidades para análise das relações entre as linhas da matriz original.

Nessa linha, optamos pelo particionamento da matriz através da aplicação da Decomposição em Valores Singulares (SVD). A escolha se justifica pelos resultados obtidos no tratamento do problema de Sistemas de Recomendação [8] [11]. Na Figura 11, o desenho esquemático mostra uma representação do particionamento.

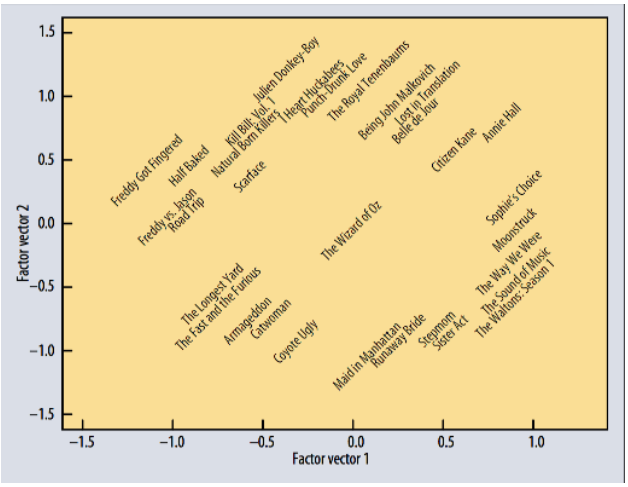

Figura 10. Mapeamento de Filmes em vetores baseado em dois fatores latentes. Extraído de Gower [6].

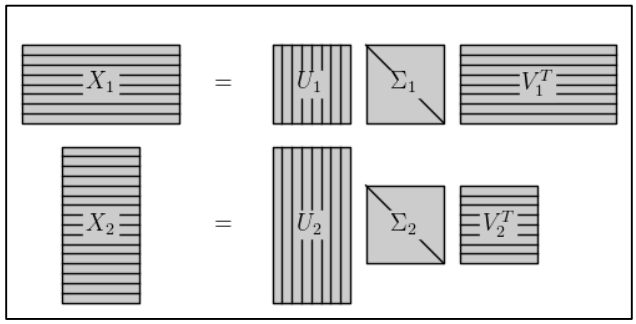

Figura 11. Particionamento da matriz SVD das matrizes X_1 e X_2, extraído de [12].

A decomposição SVD fatora uma matriz $X_{N \times M}$ nas matrizes U, Σ e V, de tal forma que a matriz Σ de vetores singulares é uma matriz quadrada, de ordem S = min(N, M). A matriz $U_{N \times S}$ acaba por armazenar em suas colunas os vetores singulares à esquerda, enquanto a matriz $V_{S \times M}$ guarda os vetores singulares à direita da matriz original.

Com a finalidade de investigar as questões propostas, a matriz utilizada foi a matriz U, que guarda informações latentes entre fatores e políticos. Após a fatoração, foram aplicados algoritmos de agrupamento sobre a matriz, como *K-Means* e *DBScan* [9], a fim de identificar afinidade entre os discursos de congressistas.

Foi realizada uma exploração espacial dos dados dos discursos, sendo este espaço formado pelos fatores obtidos por uma transformação sobre a matriz políticos-palavras. O espaço considerado caracteriza aspectos latentes que são dificilmente percebidos a partir de uma simples avaliação visual da matriz de origem. Por convenção, a transformação SVD implementada nos pacotes computacionais ordena os fatores a partir de seus respectivos autovalores, em ordem decrescente [1]. Essa vantagem acaba por discriminar ordenadamente fatores que contenham informações mais significativas primeiro, permitindo alinhar ou contrapor os dados com um número mínimo de dimensões. Para facilitar a visualização neste artigo, escolhemos representar graficamente os dois primeiros fatores, como ilustrado na Figura 12.

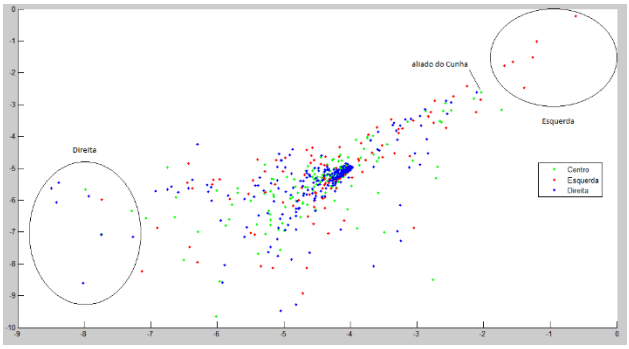

Figura 12. Gráfico de dispersão completo da matriz U, colorizado pelo critério alinhamento político-partidário, primeiro e segundo fatores.

Uma vez aplicada a decomposição na matriz normalizada, composta pelos vetores de *bag-of-word*, foram analisados os 6 primeiros fatores da matriz U, utilizando agrupamento por *K-Means* com 5 núcleos e algoritmo *DB-Scan* com ε=0,005 e mínimo de vizinhos=4, gerando os respectivos resultados apresentados nas Figura 13 Figura 14.

É possível notar que ambos os algoritmos apontaram dois grandes grupos. No entanto, é preciso introduzir informação supervisionada para analisar o significado desta informação. Inicialmente, optamos por utilizar a orientação ideológica dos partidos políticos brasileiros fornecidos pela *Wikipedia* como forma de verificar se o alinhamento ideológico influencia nesta organização. No entanto, esta definição, ainda que formalizada pelos próprios partidos, é falha na intenção de definir a prática dos parlamentares. Um exemplo seriam partidos de centro que não se auto definem centro.

Para uma melhor visualização dos resultados, optamos pela seguinte divisão apresentada na Figura 12:

- Partidos de esquerda: PCO, PSTU, PCB, PSOL, PCDOB, PT, PSB
- Partidos de direita: PR, PSD, PP, PSDB, PSC, PTC, PHS, PSDC, PROS, DEM, PSL
- Partidos de centro: demais partidos

O próximo passo foi rotular os pontos com o nome dos parlamentares, uma vez que podemos verificar alguns pontos mais afastados do cluster que reúnem algumas características particulares. Nas Figura 15 Figura 16 estão versões ampliadas das regiões circuladas na Figura 12 identificadas como regiões que contém parlamentares *outliers*, e cujo alinhamento político parece ser coeso dentro da classificação partidária proposta.

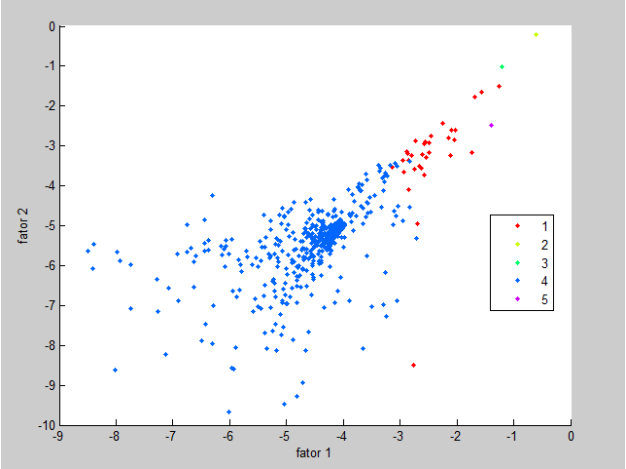

Figura 13. Aplicação do *K-Means* sobre a matriz U.

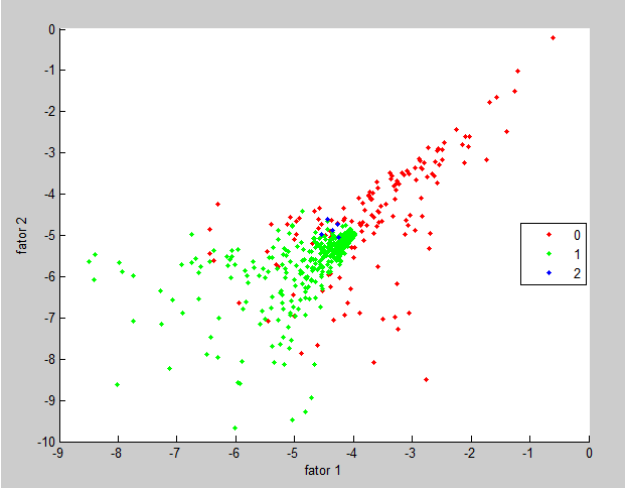

Figura 14. Aplicação do *DB-Scan* sobre a matriz U.

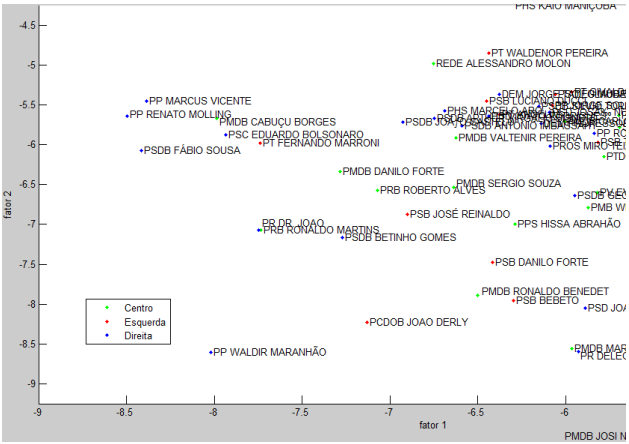

Figura 15. Detalhe inferior esquerdo, constando legenda e nomes de congressistas com valores baixos para os fatores latentes 1 e 2.

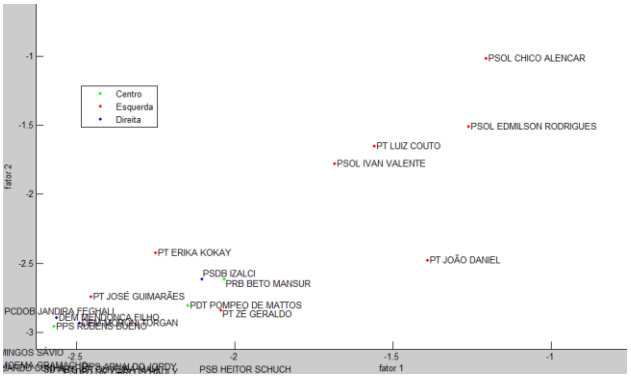

Figura 16. Detalhe superior direito, constando legenda e nomes de congressistas com valores altos para os fatores latentes 1 e 2.

4. CONCLUSÃO

O objetivo desse artigo foi realizar uma análise dos discursos de parlamentares brasileiros e investigar a similaridade entre os discursos. Além disso, foram definidas questões relevantes para análise. Considerando que a análise dos dados pode contribuir de forma a facilitar o entendimento das coligações e parcerias, esta análise pode ser útil para a identificação e comprovação de cenários.

A principal contribuição do trabalho é evidenciar a possibilidade de analisar dados textuais e encontrar tendências que podem colaborar de alguma forma para discussões e definições sócio-políticas.

A análise dos discursos permitiu identificar que há uma similaridade entre discursos do mesmo partido, assim como em partidos com ideais semelhantes. Embora não seja o objetivo deste trabalho discutir os aspectos políticos da análise, é possível concluir que alguns partidos possuem grande similaridade com outros, evidenciando parcerias e coligações. Além disso, a análise dos dados poderia ser utilizada para prospectar futuras tendências e cenários na política brasileira.

Podemos apontar como trabalhos futuros, a análise de algumas questões como: o alinhamento dos discursos de parlamentares com setores da economia e se o alinhamento acontece de forma positiva ou negativa. Além disso seria interessante analisar se há citação entre os parlamentares e partidos e se ocorre de forma positiva ou negativa.

5. REFERÊNCIAS

[1] Bruza, P. and Weeber, M. 2008. *Literature-based discovery*. Springer Science & Business Media.

[2] Dario, D. 2010. Aplicação de técnicas de visualização de informação em ferramentas para apoio à avaliação formativa em sistemas de EaD. (2010).

[3] Discursos e Notas Taquigráficas: *http://www.camara.leg.br/internet/SitaqWeb/pesquisaDiscursos.asp*. Accessed: 2017-05-15.

[4] Frey, B.J. and Dueck, D. 2007. Clustering by passing messages between data points. *science*. 315, 5814 (2007), 972–976.

[5] Gomes, L.F.O. and Tavares, J.M.R. 2011. Percepção humana na visualização de grandes volumes de dados. *Actas do 10º Congresso Iberoamericano de Engenharia Mecânica (CIBEM 10)* (2011).

[6] Gower, S. 2014. Netflix Prize and SVD. (2014).

[7] Jasser Al-Kassab, Z.M.O. 2013. Information visualization to support management decisions. *International Journal of Information Technology and Decision Making*. (2013).

[8] Koren, Y. et al. 2009. Matrix factorization techniques for recommender systems. *Computer*. 42, 8 (2009).

[9] Manning, C.D. et al. 2008. *Introduction to information retrieval*. Cambridge university press Cambridge.

[10] Orengo, V.M. and Huyck, C.R. 2001. A Stemming Algorithmm for the Portuguese Language. *Spire* (2001), 186–193.

[11] Paterek, A. 2007. Improving regularized singular value decomposition for collaborative filtering. *Proceedings of KDD cup and workshop* (2007), 5–8.

[12] Plot a visual representation of an SVD — astroML 0.2 documentation: *http://www.astroml.org/book_figures/chapter7/fig_svd_visual.html*. Accessed: 2017-05-16.

[13] Robert, C. 2014. *Machine Learning, a Probabilistic Perspective*. Taylor & Francis.

[14] Rousseeuw, P.J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*. 20, (1987), 53–65.