Relevance, diversity and serendipity in content recommendation using clustering

Fernando Henrique da Silva Costa[1], **Andrei Martins Silva**[1], **Sarajane Marques Peres**[1]

[1]Programa de Pós-Graduação em Sistemas de Informação
Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
03828-000 – São Paulo – SP – Brazil

{fhscosta0993, andreimartins, sarajane}@usp.br

Abstract. *In this paper, over-specialization in content-based recommender systems is explored through the definition and analysis of recommendation strategies aiming at quality in terms of relevance, diversity and serendipity. Clustering is applied as the basis for building these strategies, applied to the news context. The results show the feasibility of the proposed strategies.*

1. Introdução

Sistemas de recomendação (SRs) são softwares que recomendam itens adequados a usuários [Massa and Avesani 2007]. Para predizer os interesses dos usuários, os SRs usam informações sobre itens, usuários e interações entre eles [Lu et al. 2015]. Há diversos cenários para um SR, sendo de interesse deste trabalho, a recomendação de notícias.

Neste trabalho, é explorada a arquitetura de SRs baseadas em conteúdo (SRbC), que recomendam itens similares àqueles que o usuário já mostrou interesse. Segundo [Adomavicius and Tuzhilin 2005], SRs que usam estratégias de recomendação baseadas em similaridade tendem não embutir novidade ou surpresa em suas recomendações. Assim, é uma necessidade atual implementar SRs que produzam recomendações serendipitosas (que sejam relevantes e surpreendentes) [Iaquinta et al. 2008]. O objetivo deste trabalho é explorar recomendações por conteúdo que incorporem três aspectos: relevância, diversidade e serendipidade. Seis estratégias de recomendação são avaliadas qualitativamente sob esses aspectos, via interpretação feitas pelos pesquisadores-autores. As estratégias de recomendação são baseadas em agrupamentos gerados pelo algoritmo *k-means++* [Arthur and Vassilvitskii 2007] com distância cosseno [Zhong 2005]. A motivação para a escolha de agrupamento é usar como base para a recomendação as relações de similaridade natural entre as notícias. Neste artigo, é defendido que essas relações são bem caracterizadas pela similaridade entre o conteúdo completo das notícias, em oposição à análises baseadas apenas no tema central da notícia. Geralmente, o tema central da notícia é usado para categorizá-la em cadernos/canais dentro de um portal. O uso do agrupamento é uma alternativa ao uso dos cadernos/canais como base de recomendação.

As estratégias de recomendação analisadas neste artigo são: duas para produzir recomendações relevantes, i.e., que atendam ao interesse imediato de leitura do usuário; duas que atendam ao aspecto de diversidade, i.e., recomendem notícias que estejam fora das expectativas iniciais de leitura do usuário; e duas com a finalidade de apresentar recomendações serendipitosas, i.e., notícias que sejam relevantes e inesperadas ao usuário. Este artigo é divido em seis seções: seção 2 para o referencial teórico; seção 3 para trabalhos correlatos; seção 4 para apresentação das estratégias de recomendação; seção 5 para apresentação dos experimentos, resultados e análises; seção 6 para as conclusões.

2. Referencial teórico

Nesta seção são apresentados os conceitos teóricos: SRbCs, representação vetorial para dados textuais, o algoritmo *k-means++* e índices de validação de agrupamentos.

2.1. Sistemas de recomendação baseados em conteúdo

Um sistema de recomendação baseado em contéudo (SRbC) recomenda itens novos com base nas suas similaridades aos itens que foram apreciados pelo usuário [Lops et al. 2011]. Assim, o SRbC analisa um conjunto de itens avaliados pelo usuário e constrói um perfil de interesse baseado nas características desses itens [Lops et al. 2011].

Os SRbCs possuem algumas vantagens e limitações. Segundo [Lops et al. 2011], as vantagens são: **independência de outros usuários**, sendo preciso somente as avaliações do usuário para recomendar itens a ele; **transparência**, em que a lista de recomendação contém as descrições que possibilitaram a sua construção; e **novo item**, em que o SRbC é capaz de recomendar itens recém adicionados, ainda que não possuem avaliações. As limitações, segundo [Adomavicius and Tuzhilin 2005], são: **análise limitada de conteúdo**, em que o SRbC é limitado pelas características associadas aos itens; **super especialização**, em que os itens recomendados são semelhantes aos itens apreciados pelo usuário; e **novo usuário**, uma vez que ao ingressar no sistema, ele não avaliou itens.

A avaliação da qualidade de uma recomendação é realizada sob diferentes aspectos. Neste artigo, três aspectos são de interesse: relevância, diversidade e serendipidade. A serendipidade é definida em termos de relevância (tomada como acurácia neste estudo) e surpresa (tomada como diversidade neste estudo): em [Ge et al. 2010], ela é definida como uma medida do quão as recomendações são atraentes e surpreendentes ao usuário; em [Shani and Gunawardana 2011], ela é uma medida do quão surpreendente são as recomendações de sucesso (de relevância). Segundo [Herlocker et al. 2004], avaliar a serendipidade de recomendações auxilia na superação da limitação de super especialização.

A construção de SRs capazes de fazer recomendações serendipitosas requer o alcance da capacidade de recomendar itens surpreendentes e relevantes de forma equilibrada [Shani and Gunawardana 2011]. No entanto, esses dois aspectos constituem objetivos antagônicos, pois itens relevantes são os mais adequados às preferências e restrições do usuário [Ricci et al. 2011], enquanto itens surpreendentes são desconhecidos e inesperados em uma lista de recomendação pelo usuário [Shani and Gunawardana 2011].

2.2. Representação vetorial para textos

Para que dados textuais sejam analisados pelo algoritmo *k-means++*, é preciso criar uma representação vetorial para eles. Há três representações comumente usadas: binária, por frequência dos termos e a *tf-idf* [Silva et al. 2016]. Ao utilizar essas representações, uma matriz de documentos por termos é criada. Na representação binária, se o termo está presente em um documento, a célula correspondente a relação entre eles é preenchida com 1, caso contrário, é preenchida com 0. A segunda representação, conhecida como *term-frequency* (*tf*), preenche as células da matriz com o valor da ocorrência absoluta do termo no documento. Enquanto essa representação considera apenas a distribuição de frequências em um documento, a representação *tf-idf* adiciona a informação da ocorrência de um termo sobre todos documentos. A equação que calcula o valor de *idf* é $idf(t_i) = log\ n/\eta_{t_i}$, sendo t_i, um termo, n, o número total de documentos na coleção e η_{t_i}, o número

de documentos em que t_i ocorre. Sendo $tf(t_i, d_j)$, a frequência do termo t_i no documento d_j, tem-se $tfidf(t_i, d_j) = tf(t_i, d_j) * idf(t_i)$.

2.3. K-means ++ e validação de agrupamentos

O algoritmo *k-means* divide n vetores (dados em um espaço de m dimensões) em k grupos, maximizando a similaridade intragrupo e minimizando a similaridade intergrupos [Hartigan and Wong 1979, Silva et al. 2016]. No processo iterativo para implementação do *k-means* há três passos: k centroides iniciais são gerados para k grupos; cada vetor é associado ao centroide C mais similar a ele; cada centroide é atualizado para ser o vetor médio dos vetores a ele associados. Os dois últimos passos se repetem até estabilizar os centroides. Formalmente, seja um conjunto de dados $X = \{\overrightarrow{x_i}\}$, com $i = 1, ..., n$, a ser agrupado em k grupos $G = \{g_j\}, j = 1, ..., k$. O algoritmo busca por uma partição em X que minimize o problema $P(U, G) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ji} dist(\overrightarrow{g_j}, \overrightarrow{x_i})$, em que $\overrightarrow{g}_j$ é o vetor centroide em g_j, $U = \{0, 1\}^{k,n}$ é a matriz indicadora para associação dos n vetores aos k grupos e $dist$ é, neste trabalho, a distância cosseno. O algoritmo *k-means++* é uma extensão do *k-means* em que é escolhido um dado $\overrightarrow{x} \in X$ aleatoriamente como o primeiro centroide e os próximos centroides são escolhidos pela seleção de $\overrightarrow{x_l} \in X$ com probabilidade $dist_{min}(\overrightarrow{x_l}) \ / \sum_{\overrightarrow{x} \in X} dist_{min}(\overrightarrow{x})$, sendo $dist_{min}(\overrightarrow{x})$, a menor distância de um dado $\overrightarrow{x} \in X$ para o centroide mais próximo já escolhido [Arthur and Vassilvitskii 2007].

Para validar os agrupamentos, índices baseados em critérios externos e internos foram utilizados: Informação Mútua Normalizada (NMI) [Strehl and Ghosh 2002] como índice externo; Silhouette [Rousseeuw 1987] como índice interno.

3. Trabalhos correlatos

Em [Bobadilla et al. 2013], os autores apresentam uma revisão sobre SRs, discutindo a taxonomia, arquiteturas, algoritmos de recomendação, conjunto de dados, medidas de avaliação e tendências. Em [Kotkov et al. 2016], os autores discutem análise de serendipidade, apresentando definições, algoritmos de recomendação e estratégias de avaliação.

Estratégias para embutir e avaliar surpresa e serendipidade em recomendações são estudadas em [Zheng and Ip 2012, Jenders et al. 2015, Piao and Whittle 2011, Xiao et al. 2014]. Em [Zheng and Ip 2012], um *framework* foi desenvolvido para equilibrar os graus de surpresa de recomendações. Medidas de relevância e de diversidade são combinadas para explorar áreas de potencial interesse dos usuários e recomendar filmes. Em [Jenders et al. 2015], os autores discutem serendipidade em SRs para notícias. As recomendações são feitas com base na similaridade de conteúdo das notícias e um experimento com usuários é realizado para avaliar o quão serendipitosas elas são. Em [Piao and Whittle 2011], com uso de processamento de língua natural (NLP), os autores extraem os interesses dos usuários e sugerem conexões serendipitosas. Em [Xiao et al. 2014], os autores propõem um *framework* de recomendações serendipitosas de artigos acadêmicos e duas medidas que avaliam tal característica.

Alguns trabalhos resolvem a tarefa de agrupamento com a finalidade de embasar estratégias de recomendações. Em [Liao and Lee 2016], os autores aplicam *self-constructing clustering (SCC)* para agrupar itens e usuários, e sobre isso constroem recomendações. Em [Gong 2010], os autores propõem um abordagem de recomendação

personalizada que une agrupamentos de usuários, com base em suas avaliações sobre os itens, e agrupamento de itens, com base nas avaliações recebidas.

4. Estratégias de recomendação

Neste artigo é apresentada uma análise qualitativa sobre recomendações relevantes, diversas e serendipitosas, mediante a aplicação de seis estratégias de recomendação. Cada estratégia é, presumidamente, mais adequada para realizar recomendações com determinada característica. As estratégias são aplicadas sobre um conjunto de notícias para o qual é pressuposto existir uma partição. A construção da partição, neste artigo, é obtida com o algoritmo *k-means++* e a distância cosseno. A partir dos k grupos gerados, estratégias de recomendação são aplicadas para criar uma lista L de notícias que alcance um de três aspectos: relevância, diversidade e serendipidade. A base da recomendação é uma notícia semente (xs) que é acessada pelo usuário que receberá a recomendação. O acesso à notícia semente é o que caracteriza o perfil do usuário, mitigando o problema de "novo usuário". Nesse contexto, a análise das características da recomendação assume que: (a) o sistema não conhece outras preferências do usuário, a não ser a preferência pela notícia xs; (b) o usuário não teve acesso prévio a nenhuma das notícias que está na lista L; e (c) a recomendação é baseada no texto das notícias e na sua organização em grupos.

4.1. Estratégias para recomendações relevantes

As duas primeiras estratégias de recomendação são projetadas para gerar uma lista de recomendação de notícias relevantes, i.e., uma lista com notícias que estejam no hábito de leitura do usuário, considerando que esse hábito é bem representado pela notícia semente e que, portanto, quanto mais similar são as notícias (semente e recomendadas), mais relevante é a recomendação. Essas estratégias assumem que $L = \{xo_1, xo_2, \ldots, xo_l\} \quad | \quad xo_i \sim xs, i = \{1, \ldots, l\}$, em que xo_i é uma notícia no conjunto de notícias disponíveis, l é a quantidade de notícias em L e $a \sim b$ significa que uma notícia a é similar à notícia b. As duas estratégias são definidas da seguinte forma:

- recomendar as notícias mais similares à notícia semente e que sejam pertencentes ao mesmo grupo em que a notícia semente, i.e.: $\arg\max_i \ (xo_i \sim xs) \ \forall \ xo_i \in L, \ i = \{1, \ldots, l\} \mid xo_i, xs \in g_j$ e $xo_i \neq xs$;
- recomendar aleatoriamente notícias pertencentes ao mesmo grupo da notícia semente, i.e: $rand(xo_i) \ \forall \ xo_i \in L, \ i = \{1, \ldots, l\} \mid xo_i \ \wedge \ xs \in g_j$ e $xo_i \neq xs$.

A figura 1 ilustra as estratégias de recomendação descritas. A notícia semente está em azul, enquanto em amarelo estão as notícias que compõem a lista de recomendação. Nesta figura, o aglomerado de documentos representa um único grupo.

(a) (b)

Figura 1. Estratégias para recomendações relevantes

4.2. Estratégias para recomendações diversificadas

As estratégias para recomendar notícias diversificadas são projetadas no sentido de causar surpresa. De acordo com [Shani and Gunawardana 2011], a diversidade é definida como o oposto de similaridade. Em alguns casos, recomendar uma lista de itens similares aos que foram apreciados pelo usuário pode não ser interessante, ou mesmo útil, para ele. Essas duas estratégias visam gerar listas de recomendações que tragam diversidade de assuntos para o contexto do usuário, e assumem que $L = \{xo_1, xo_2, \ldots, xo_l\} \mid xo_i \neq xs, i = \{1, \ldots, l\}$, em que xo_i e l são como definido anteriormente e $a \neq b$ significa que uma notícia a possui conteúdo diferente em relação ao conteúdo da notícia b. As duas estratégias para recomendações diversificadas são definidas da seguinte forma:

- recomendar notícias aleatoriamente escolhidas dentro do grupo mais distante do grupo ao qual a notícia semente pertence. i.e.[1]: $rand(xo_i) \ \forall \ xo_i \in L, \ i = \{1, \ldots, l\} \mid xo_i \in g_j, xs \in g_o, \ j, o \in \{1, k\}$ e $\arg\max_j(dist(\overrightarrow{g_j}, \overrightarrow{g_o}))$;
- recomendar notícias aleatoriamente escolhidas dentro de todos os grupos exceto daquele ao qual a notícia semente pertence, i.e.: $rand(xo_i) \ \forall \ xo_i \in L, \ i = \{1, \ldots, l\} \mid xo_i \in g_j, \ xs \in g_o, \ j = \{1, \ldots, k\}, \ o \in \{1, k\}$ e $j \neq o$.

A figuras 2a e 2b ilustram as estratégias para recomendações diversificadas. As notícias estão agrupadas em k grupos. A notícia semente está em azul e pertence ao primeiro grupo, enquanto as notícias que formam a lista de recomendação estão em amarelo e: (a) pertencem ao grupo mais distante do grupo da notícia semente (grupo k na figura); (b) pertencem a grupos distintos do grupo ao qual a notícia semente pertence.

4.3. Estratégia para recomendações serendipitosas

As duas últimas estratégias de recomendação tem como finalidade gerar listas de recomendação em que as notícias sejam serendipitosas. Para isso, as estratégias são definidas de forma a combinar elementos das estratégias já apresentadas, em conformidade com as definições de serendipidade apresentadas em [Shani and Gunawardana 2011, Ge et al. 2010], e assumem que $L = \{xo_1, xo_2, \ldots, xo_l\} \mid xo_i \asymp xs, i = \{1, \ldots, l\}$, em que no_i e l são como definido anteriormente e $a \asymp b$ significa que a noticia a possui alguma similaridade com relação à notícia b, porém traz algum conteúdo novo (diferente). As estratégias para alcance de serendipidade são definidas da seguinte forma:

- recomendar as notícias mais similares à notícia semente, considerando notícias pertencentes ao grupo mais próximo do grupo ao qual pertence a notícia semente, i.e.: $\arg\max_i \ (xo_i \sim xs) \ \forall \ xo_i \in L, \ i = \{1, \ldots, l\} \mid xo_i \in g_j, \ xs \in g_o$ e $\arg\min_j(dist(\overrightarrow{g_j}, \overrightarrow{g_o})), \ j, o \in \{1, k\}$;
- recomendar as notícias mais similares à notícia semente, considerando notícias pertencentes ao grupo que representa um balanceamento entre o grupo mais próximo e o grupo mais distante do grupo ao qual pertence a notícia semente, i.e.: $\arg\max_i \ (xo_i \sim xs) \ \forall \ xo_i \in L, \ i = \{1, \ldots, l\} \mid xo_i \in g_j, \ xs \in g_o$ e $\arg\min_j(dist(\overrightarrow{g_j}, \overrightarrow{\mathbf{g}})), \ j \in \{1, k\}$, em que $\overrightarrow{\mathbf{g}} \approx \frac{\sum_{j=1 \mid j \neq o}^{k} \overrightarrow{g_j}}{k-1}$.

As figuras 2c e 2d ilustram as estratégias para recomendações serendipitosas. A partição encontrada possui k grupos. Considere que o grupo 2 é o mais próximo do grupo

[1]Para o cálculo da distância entre grupos, considere a distância entre os seus centroides.

1, o grupo 3 é o mais próximo do grupo 2, e assim por diante. Se a notícia semente pertence ao grupo 1 então, a primeira estratégia para alcance da serendipidade recomenda notícias do grupo 2, enquanto a segunda recomenda notícias do grupo i. Na figura, a notícia semente está em azul e as notícias recomendadas estão em amarelo.

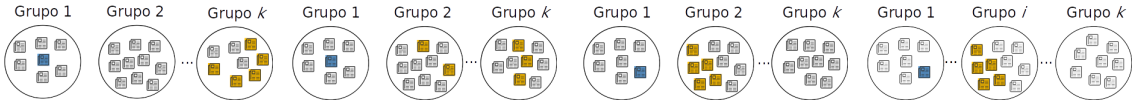

Figura 2. Estratégias de recomendação diversificadas (a,b) e serendipitosas (c,d)

5. Experimentos, resultados e análises

Para a realização dos experimentos, foi utilizado o algoritmo *k-means++* com distância cosseno implementado no ambiente MATLAB[2]. Os experimentos foram realizados considerando dois valores para k: *Exp.* $\#1$ ($k = 10$), com notícias organizadas em dez categorias (cf. seção 5.1) de forma que uma partição com dez grupos pudesse ser avaliada usando um índice de validação de agrupamento externo e considerando tais categorias como *verdade fundamental*; *Exp.* $\#2$ ($k = 7$), com a união de categorias e com um número de grupos reduzido para verificar se resultados diferenciados seriam obtidos.

Dez execuções do *k-means++* foram realizadas em 30 conjuntos de dados gerados a partir do conjunto de notícias. As execuções foram avaliadas pelos índices de validação de agrupamentos baseados em critérios externos (NMI) e internos (Silhouette). Para cada índice foram calculados as médias, os desvios padrões e os coeficientes de variação para ilustrar a robustez do algoritmo às variações das condições iniciais. O uso de índices de validação são importantes para guiar a escolha de uma boa partição para ser usada como base de recomendação. No restante da seção são apresentados o conjunto de dados, as etapas de pré-processamento realizadas sobre o conjunto de dados, os resultados alcançados em cada experimento e a análise qualitativa sobre as características das recomendações.

5.1. Conjunto de dados

O conjunto de dados *Corpus EBC* usado nos experimentos é composto por 1.157 notícias do portal de notícias Empresa Brasil de Comunicação (EBC)[3]. Cada notícia pertence a um dos cadernos encontrados no portal, sendo eles: *Cidadania*, *Cultura*, *Educação*, *Esportes*, *Infantil*, *Política*, *Economia*, *Internacional*, *Tecnologia* e *Geral*. As notícias são associadas aos cadernos de acordo com a interpretação dos produtores de conteúdo do portal. O caderno *Geral* diz respeito a notícias diversificadas que, sob o ponto de vista dos produtores de conteúdo, não são aderentes aos demais cadernos. Assim, trata-se de um caderno com notícias que versam sobre assuntos não necessariamente relacionados.

Como forma de estabelecer uma verdade fundamental para a categorização das notícias, cada uma delas foi associado um rótulo correspondente ao caderno no qual ela foi publicada no portal. Assim, dez rótulos foram usados e a distribuição (absoluta e relativa) das notícias nas categorias é listada na tabela 1. Como pode ser observado na tabela, trata-se de um conjunto de notícias com distribuição de categorias desbalanceada.

[2]https://www.mathworks.com/products/matlab.html
[3]http://www.ebc.com.br/

Tabela 1. Conjunto de notícias: *Corpus EBC*

Cadernos	Rótulo	# de notícias	%	Cadernos	Rótulo	# de notícias	%
Educação	3	257	22,2	Cidadania	1	96	8,3
Esportes	4	200	17,3	Internacional	9	89	7,7
Tecnologia	7	125	10,8	Política	6	86	7,5
Geral	10	125	10,8	Economia	8	41	3,5
Cultura	2	124	10,7	Infantil	5	14	1,2

5.2. Pré-processamento

Neste trabalho, foi usada a abordagem de *bag-of-words* [Salton et al. 1975], que transforma uma coleção de textos em uma matriz $X^{n \times m}$, em que cada linha representa um documento i, com $i = \{1, \ldots, n\}$, cada coluna representa um termo j, com $j = \{1, \ldots, m\}$ e cada elemento x_{ij} indica a ocorrência de um termo j no documento i. Neste trabalho, foram geradas 30 matrizes, $X_1, X_2, X_3, \ldots, X_{30}$, a partir de 30 combinações de operações de pré-processamento. As operações aplicadas foram: *tokenização* com remoção de números, símbolos e pontuações; remoção de *stopwords*[4] (opcional); seleção de termos relevantes (por frequência mínima ou por *tf-idf*); e mapeamento dos textos para uma representação vetorial usando representação binária, *tf* ou *tf-idf*. A seleção de termos relevantes foi feita: por frequência mínima, mantendo apenas os termos que apareceram em ao menos d documentos da coleção; por *tf-idf*, selecionando os q termos com maior escore *tf-idf*. Neste trabalho foram utilizados $d = 2$ e $q = \{5, 10, 15, 20\}$. Por fim, foi aplicada normalização euclidiana nas matrizes, com exceção às matrizes binárias.

5.3. Resultados

Esta seção apresenta os resultados da execução do *k-means++* sobre as matrizes de dados e uma análise qualitativa sobre características de relevância, diversidade e serendipidade das listas de recomendação L. Listas L de tamanho $l = 3$ foram obtidas a partir da aplicação das estratégias de recomendação descritas na seção 4 sobre os cinco agrupamentos de maior qualidade, segundo os índices de validação, obtidos em cada experimento. Para escolha dos melhores agrupamentos, foram utilizados os valores de média de I_{NMI} e I_{SIL}.

Exp. #1: A tabela 2 apresenta os valores da validação dos cinco melhores agrupamentos com $k = 10$, e as opções de pré-processamento que criaram as matrizes nas quais esses resultados foram obtidos. Os resultados são apresentados como média, desvio padrão (DP) e coeficiente de variação (CV), calculados sobre dez execuções. Listas de recomendação obtidas sobre grupos da matriz X_{10} foram analisadas. Essa partição foi escolhida por alcançar qualidade superior na validação externa, uma vez que sob a validação interna, as cinco melhores partições apresentam qualidade semelhante. As notícias são apresentadas em termos de seus títulos. A notícia semente usada para gerar recomendações foi escolhida aleatoriamente. Ela pertence ao caderno *Cidadania* e é intitulada: *"Polícia investiga agressões a LGBTs em show no Rio"*.

A lista gerada pela **primeira estratégia de recomendação visando a relevância** contém as notícias: *"Ministro quer ampliar treinamento de policiais brasileiros com oficiais dos EUA"*, *"Idosa é agredida a pedradas no RJ e família denuncia intolerância religiosa"* e *"Desligamento de funcionários do projeto Viver preocupa vítimas de abuso*

[4]Lista de *stopwords* disponível em http://snowball.tartarus.org/algorithms/portuguese/stop.txt.

Tabela 2. Qualidade de agrupamento (Exp. #1), de acordo com índices de validação e opções de pré-processamento. Melhores resultados em negrito

Matriz	I_{NMI}			I_{SIL}			Operações de pré-processamento			
	Média	DP	CV	Média	DP	CV	*Stopw*	Atrib	Selec	Repres
X_2	0,56	0,02	0,04	0,08	0,008	0,01	sim	freq. min	2	bin
X_4	0,51	0,01	0,02	0,08	0,004	0,009	sim	*tf-idf*	5	bin
X_6	0,51	0,01	0,02	**0,08**	**0,003**	**0,007**	sim	*tf-idf*	10	bin
X_8	0,56	0,03	0,06	0,08	0,004	0,007	sim	*tf-idf*	15	bin
X_{10}	**0,57**	**0,02**	**0,03**	0,08	0,005	0,009	sim	*tf-idf*	20	bin

sexual". A primeira recomendação pertence ao caderno *Geral* e as demais pertencem ao mesmo caderno da notícia semente. Existe similaridade entre as notícias recomendadas e a notícia semente, pois todas tratam de violência e segurança pública. As duas últimas notícias tratam também de violência e minorias, assim como a notícia semente. A **segunda estratégia visando relevância** gerou a lista com as notícias: *"Auxiliar de creche é baleada dentro de escola no Rio"*, *"Estado do Rio ganha força-tarefa social para atuar na redução da violência"* e *"Relatores da ONU criticam "ataques"aos direitos ambientais e indígenas no país"*. Todas as notícias pertencem ao caderno *Cidadania*, o mesmo da notícia semente. As duas primeiras são altamente similares à notícia semente, pois dissertam sobre a violência e segurança pública no Rio de Janeiro. A terceira, versa sobre um grupo social diferente (indígenas), porém apresenta elementos sobre um tipo de violência.

A lista gerada pela **primeira estratégia de recomendação visando diversidade** contém as as notícias: *"No Mundo da Bola analisa amistoso entre Brasil e Argentina"*, *"Atlético-PR se prepara para o jogo contra o Flamengo pela Libertadores"* e *"Chapecoense lidera o Brasileirão depois de quatro rodadas"*. As notícias da lista pertencem ao caderno *Esportes*. Elas versam sobre um assunto diferente da notícia semente, portanto, apresentam diversidade em relação a ela. No entanto, todas elas versam sobre um mesmo tema, gerando uma lista sem diversidade interna. A **segunda estratégia que visa a diversidade** gerou a lista com as notícias: *"Botafogo empata com o Sport e se garante nas quartas da Copa do Brasil"*, *"No Mundo da Bola destaca a preparação da Seleção Brasileira em Porto Alegre"* e *"Justiça manda soltar militantes do MTST presos durante greve geral"*. As duas primeiras notícias pertencem ao caderno *Esportes* e a última está presente em *Geral*. Diferente da estratégia anterior, essa apresenta uma lista com diversidade interna, uma vez que há uma notícia que trata sobre um assunto diferente das outras.

A lista gerada pela **primeira estratégia visando serendipidade** contém as notícias: *"Apenas 1% dos brasileiros com deficiência está no mercado de trabalho"*, *"Projetos sociais incentivam negócios de impacto no Brasil"* e *"Instituto Vital Brazil desenvolve remédio inédito contra veneno de abelha"*. As notícias da lista pertencem aos cadernos de *Cidadania*, *Educação* e *Tecnologia*. É possível encontrar diversidade quando comparada com a notícia semente, e ao analisar o conteúdo completo das notícias, tópicos em comum entre elas são encontrados. A primeira recomendação trata de um grupo minoritário (deficientes), igualmente à notícia semente (LGBTs). A segunda disserta sobre projetos sociais que incluem o estado do Rio de Janeiro e grupos de minorias. A terceira tem como similaridade o Instituto Vital Brazil, ligado à secretaria de saúde do Rio de Janeiro. A **segunda estratégia visando a serendipidade** gerou a lista de notícias: *"Servidores e estudantes da UERJ fazem ato em defesa da instituição"*, *"Em crise, UERJ faz vestibular*

com menos da metade dos candidatos do ano passado" e *"Educadores pedem "blindagem social" em escola onde morreu Maria Eduarda"*. Todas notícias recomendadas pertencem ao caderno *Educação*. Ainda que discutam temas diferentes, apresentam similaridades com a notícia semente. As duas primeiras tratam sobre evento na UERJ-Rio de Janeiro. A notícia envolve violência dentro do âmbito escolar também no Rio de Janeiro.

Exp. #2: Na tabela 3 estão os valores dos índices de validação dos cinco melhores agrupamentos com $k = 7$ e as opções de pré-processamento que criaram as matrizes nas quais eles foram obtidos. As listas de recomendação foram obtidas sobre o agrupamento da matriz X_2, escolhido sob a mesma estratégia usada no Exp. #1. A notícia semente sorteada é do caderno *Economia* e é intitulada: *"Produtor pode solicitar recursos do Plano Safra 2017-2018 a partir de segunda"*.

Tabela 3. Qualidade de agrupamento (Exp. #2), de acordo com índices de validação e opções de pré-processamento. Melhores resultados em negrito

Matriz	I_{NMI}			I_{SIL}			Operações de pré-processamento			
	Média	DP	CV	Média	DP	CV	*Stopw*	Atrib	Selec	Repres
X_2	**0,56**	**0,02**	**0,05**	0,07	0,008	0,01	sim	freq. min	2	bin
X_6	0,45	0,03	0,08	**0,08**	**0,001**	**0,003**	sim	*tf-idf*	10	bin
X_8	0,55	0,02	0,04	0,08	0,004	0,007	sim	*tf-idf*	15	bin
X_{10}	0,54	0,02	0,04	0,08	0,004	0,008	sim	*tf-idf*	20	bin
X_{24}	0,50	0,02	0,05	0,02	0,003	0,007	sim	*tf-idf*	5	*tf-idf*

A lista gerada pela **primeira estratégia de relevância** contém as notícias: *"Privatizações devem ajudar no cumprimento da meta fiscal, diz Meirelles"*, *"Superintendente do BNDES diz que "pior já passou" para a economia"* e *"Pezão diz que Rio de Janeiro deve regularizar salários de servidores em agosto"*. Todas essas notícias pertencem ao caderno *Economia*. Elas possuem similaridades com a notícia semente, uma vez que versam sobre a economia nacional em um período. A lista gerada pela **segunda estratégia de recomendação de relevância** contém as notícias: *"TRF garante rematrícula de universitária que conseguiu vaga por meio de cota"*, *"MPF pede que empresários reparem danos ambientais causados em Angra dos Reis"* e *"Prefeitura do Rio inicia nesta terça reordenação de camelôs nas ruas"*. A primeira recomendação pertence ao caderno *Educação*, enquanto as demais estão em *Geral*. Neste caso, a aleatoriedade na escolha das notícias dentro do grupo inseriu diversidade na lista de recomendação. Porém, a presença de instituições públicas nas notícias é o fator que, provavelmente, as organizou em um mesmo grupo (diferentemente da organização original das notícias, a qual as colocou em cadernos diferentes, incluindo o caderno cuja temática é indefinida - *Geral*).

A lista gerada pela **primeira estratégia visando diversidade** contém as notícias: *"Flamengo e Cruzeiro empatam no primeiro jogo da final da Copa do Brasil"*, *"Bate Bola Nacional destaca a contratação de Éverton Ribeiro pelo Flamengo"* e *"Brasil vence a Austrália de goleada em Melbourne"*. Essas recomendações pertencem ao caderno *Esportes*. O comportamento é o mesmo observado no *Exp. #1*, inclusive o grupo com tema "esporte" foi novamente definido como o grupo de maior dissimilaridade ao grupo da notícia semente. As notícias da lista formada pela **segunda estratégia visando a diversidade** são: *"UEA Tabatinga-AM vai realizar terceira Edição dos Jogos Universitários Intercurso"*, *"Governo de Cuba critica "pressão dos EUA"para mudanças de sistema na ilha"* e *"Escolas públicas do DF normatizam uso de celular em sala de aula"*. Essas

notícias pertencem aos cadernos *Esportes*, *Internacional* e *Educação*, respectivamente. Há diversidade entre os assuntos tratados pelas recomendações e notícia semente.

A lista gerada pela **primeira estratégia que visa serendipidade** contém as notícias: *"Produção brasileira de cana-de-açúcar pode chegar a 646 milhões de toneladas"*, *"Em dez anos, Brasil deve ultrapassar os EUA na produção de soja"* e *"Aumento de geração de energia por consumidor pode mudar perfil de distribuidoras"*. As duas primeiras pertencem ao caderno *Economia*, enquanto a última encontra-se em *Tecnologia*. As duas primeiras apresentam relevância alta já que o conteúdo é relacionado à economia no setor rural, porém estão organizadas em um grupo diferente ao da notícia semente. A terceira é diferente, pois não versa sobre agricultura, porém está relacionada ao contexto econômico brasileiro. A lista formada pela **segunda estratégia que visa a serendipidade** contém as notícias: *"História Hoje: Há 48 anos, morria o poeta Guilherme de Almeida"*, *"Divulgada programação do 50º Festival de Brasília do Cinema Brasileiro, que começa em 15 de setembro"* e *"Festival de cinema LGBT exibe mais de 80 filmes no Rio até dia 16"*. As recomendações pertencem ao caderno *Cultura*. A lista apresenta notícias com diversidade comparada a semente, porém versando somente sobre assuntos de cultura nacional.

5.4. Análises

A qualidade dos agrupamentos no *Exp. #2* é levemente pior do que a obtida no *Exp. #1*, em termos de índices internos e de sensibilidade às condições iniciais. Porém, ambos experimentos revelam o mesmo comportamento em termos de discordância em relação à verdade fundamental (índice externo) e em termos de aderência dos dados aos grupos (índice interno). Os valores para o I_{NMI} mostram que há discordância em relação à organização das notícias em grupos e à categorização delas em cadernos. Essa discordância é um indício que a categorização das notícias: (a) não representa os reais relacionamentos entre elas; (b) é inconsistente, no sentido que a interpretação dos produtores de conteúdo sobre os relacionamentos entre notícias, e entre notícias e cadernos, pode influenciar, de maneiras diferentes, a decisão sobre a associação de notícias a cadernos. Os valores de I_{SIL} mostram que os agrupamentos não são livres de crítica, já que o valor próximo a 0 mostra que os dados colocados em um grupo são também aderentes ao outro grupo mais próximo dele. Este fato revela a complexidade de análise oferecida pelo conteúdo do *Corpus EBC*. O comportamento observado nos experimentos em relação a ambos os índices indicam que a escolha de uma boa partição como base de recomendação pode ser feita sobre índices internos, não sendo necessário usar a verdade fundamental.

As estratégias de recomendação baseadas em **relevância** assumem que a similaridade entre os conteúdos das notícias é importante. As estratégias atenderam à essa premissa nos experimentos. Porém, a segunda estratégia baseada em relevância considera a aleatoriedade que, em ambos os experimentos, inseriu diversidade às listas de recomendação. Esse fato revela que tal estratégia pode ser considerada quando o aspecto serendipidade está sendo buscado. Para o caso das estratégias que visaram **diversidade**, os resultados não corresponderam às expectativas inciais. Embora as notícias recomendadas versassem sobre temas diferentes da notícia semente, a primeira estratégia produziu listas sem diversidade interna, i.e., com notícias muito similares. De fato, tal estratégia tenderia a esse comportamento já que é restrita a recomendar notícias de um mesmo grupo. As estratégias que visaram **serendipidade** produziram resultados qualitativamente difíceis de analisar. No *Exp. #1*, ambas estratégias foram capazes de combinar relevância e di-

versidade, com mais ou menos ênfase a cada um desses aspectos, a depender do ponto de vista em que se analisa as notícias. Já no *Exp. #2*, a primeira estratégia gerou uma lista com forte aspecto de relevância, e a segunda estratégia gerou uma lista com forte aspecto de diversidade. Comparando os dois experimentos, nota-se que no primeiro, os resultados para recomendações relevantes são mais aderentes ao perfil do usuário, e que a serendipidade é mais facilmente percebida nas recomendações que usaram estratégias para esse fim. A queda de desempenho no *Exp. #2* pode ser devido ao número menor de grupos, pois os grupos continham notícias menos aderentes (vide índice I_{SIL}).

A última análise versa sobre a presença de notícias de cadernos diferentes em listas geradas usando similaridade (não aleatório). No *Exp. #1*, uma notícia de *Geral* é recomendada como relevante em relação a uma notícia de *Cidadania*. Notícias de três cadernos (Cidadania, Educação e Tecnologia) aparecem em uma lista gerada sobre um mesmo grupo. Situações similares ocorrem no *Exp. #2* para notícias de: (a) *Economia*, *Educação*, *Geral*; (b) *Economia* e *Tecnologia*. Esses fatos mostram a viabilidade das estratégias de agrupamento como base para recomendação como uma alternativa: (a) ao uso da base em cadernos/canais; (b) se uma organização prévia de notícias não é disponível.

6. Conclusão

Este trabalho introduziu seis estratégias de recomendação baseadas no agrupamento de notícias. Análise qualitativas verificaram o atendimento de aspectos de relevância, diversidade e serendipidade nas listas de recomendação, e mostraram a viabilidade do uso do agrupamento como base de recomendação e das estratégias de recomendação exploradas. Os próximos passos desta pesquisa são: experimentos de avaliação com usuários; análises quantitativas com medidas de surpresa e serendipidade; e a extensão das estratégias com uso de perfis históricos de usuários, co-agrupamento e de *ensemble* de agrupamentos.

Referências

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowledge and Data Eng.*, 17(6):734–749.

Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *Proc. of the 18th annual ACM-SIAM symp. on discrete algorithms*, pages 1027–1035. Soc. for Ind. and Applied Math.

Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Knowl.-based Sys.*, 46:109–132.

Ge, M., Delgado-Battenfeld, C., and Jannach, D. (2010). Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proc. of the 4th ACM Conf. on Recommender Sys.*, pages 257–260. ACM.

Gong, S. (2010). A collaborative filtering recommendation algorithm based on user clustering and item clustering. *J. of Software*, 5(7):745–752.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: a k-means clustering algorithm. *J. of the Royal Statistical Soc. - Series C (Applied Statistics)*, 28(1):100–108.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. on Information Sys.*, 22(1):5–53.

Iaquinta, L., De Gemmis, M., Lops, P., Semeraro, G., Filannino, M., and Molino, P. (2008). Introducing serendipity in a content-based recommender system. In *8th Int. Conf. on Hybrid Intell. Sys.*, pages 168–173.

Jenders, M., Lindhauer, T., Kasneci, G., Krestel, R., and Naumann, F. (2015). A serendipity model for news recommendation. In *Joint German/Austrian Conf. on Artificial Intell.*, pages 111–123. Springer.

Kotkov, D., Wang, S., and Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowl.-based Sys.*, 111:180–192.

Liao, C.-L. and Lee, S.-J. (2016). A clustering based approach to improving the efficiency of collaborative filtering recommendation. *Electronic Commerce Research and Applications*, 18:1–9.

Lops, P., De Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: state of the art and trends. In *Recommender Sys. Handbook*, pages 73–105. Springer.

Lu, J., Wu, D., Mao, M., Wang, W., and Zhang, G. (2015). Recommender system application developments: a survey. *Decision Support Sys.*, 74:12–32.

Massa, P. and Avesani, P. (2007). Trust-aware recommender systems. In *Proc. of the 2007 ACM Conf. on Recommender Sys.*, pages 17–24. ACM.

Piao, S. and Whittle, J. (2011). A feasibility study on extracting twitter users' interests using nlp tools for serendipitous connections. In *IEEE 3rd Int. Conf. on Social Comp. and IEEE 3rd Int. Conf. on Privacy, Security, Risk and Trust*, pages 910–915. IEEE.

Ricci, F., Rokach, L., and Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. of Comp. and Applied Math.*, 20:53–65.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Comm. of the ACM*, 18(11):613–620.

Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer.

Silva, L. A., Peres, S. M., and Boscarioli, C. (2016). *Introdução à mineração de dados: com Aplicações em R*. Elsevier.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. of Machine Learning Research*, 3(Dec):583–617.

Xiao, Z., Che, F., Miao, E., and Lu, M. (2014). Increasing serendipity of recommender system with ranking topic model. *Applied Math. & Information Sciences*, 8(4):2041.

Zheng, Q. and Ip, H. H. (2012). Customizable surprising recommendation based on the tradeoff between genre difference and genre similarity. In *IEEE/WIC/ACM Int. Conf. on WEB Intell. and Intell. Agent Technology*, volume 1, pages 702–709. IEEE.

Zhong, S. (2005). Efficient online spherical k-means clustering. In *Proc. Int. Joint Conf. on Neural Networks*, volume 5, pages 3180–3185. IEEE.