Agrupamento Hierárquico e Multivisão de Eventos por meio de Grafos de Consistência

Paulo H. L. de Paula[1], Westerley S. Reis[2], Solange O. Rezende[3], Ricardo M. Marcacini[1]

[1] Universidade Federal de Mato Grosso do Sul
Câmpus de Três Lagoas (CPTL)
paulo.paula@alunos.ufms.br, ricardo.marcacini@ufms.br
[2] Universidade Federal de Mato Grosso do Sul
Faculdade de Computação (FACOM)
westerreis@gmail.com
[3] Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação (ICMC)
solange@icmc.usp.br

Abstract. A análise de eventos tem recebido atenção recentemente devido à popularização de plataformas web para publicação de conteúdo, especialmente portais de notícias, redes sociais, blogs e fóruns. Essas plataformas armazenam eventos por meio de textos a respeito de diversos setores da sociedade e podem ser vistas como uma representação digital (mundo virtual) dos eventos que ocorrem em nosso mundo real. Assim, agrupamento de eventos é uma tarefa importante para organizar e mapear os eventos desse mundo virtual para nosso mundo físico, o que permite a realização de diversos estudos sociais, políticos e econômicos. Nesse trabalho é apresentada uma abordagem para agrupamento hierárquico e multivisão de eventos extraídos de textos. As diferentes informações sobre os eventos, como informação textual, informação temporal e informação geográfica são consideradas diferentes visões durante a tarefa de agrupamento. Enquanto as abordagens existentes exigem que o usuário defina parâmetros sobre como utilizar informação temporal e geográfica no agrupamento de eventos, a abordagem proposta permite aprender automaticamente restrições de tempo e local. Para tal, foi proposta uma estrutura denominada grafo de consistência que representa o consenso de agrupamentos entre as diferentes visões. Uma avaliação experimental com oito conjuntos de eventos de *benchmark* revelou que a abordagem proposta é superior à abordagem tradicionalmente utilizada na área, apresentando ainda o diferencial de permitir a visualização das relações entre eventos por meio do grafo de consistência.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: análise de eventos, agrupamento multivisão, grafo de consistência

1. INTRODUÇÃO

A análise de eventos é uma tarefa útil para estudar fenômenos importantes que ocorrem em locais específicos e em um determinado período de tempo [Hogenboom et al. 2016]. Diversos estudos sociais, políticos e econômicos são beneficiados a partir de pesquisas envolvendo análise de eventos [Florence et al. 2017], como monitoramento de conflitos urbanos, análise de epidemias, estudo de efeitos climáticos, análise de opinião e sentimentos, análise de tendências econômicas, bem como a construção de indicadores inteligentes em diversos domínios, a exemplo de agronegócios e medicina. Embora a análise de eventos tenha sido mais frequentemente empregada para fenômenos que ocorram em locais georreferenciados (i.e., com latitude e longitude), a análise de eventos pode ser aplicada em qualquer problema em que se espera identificar causalidade entre eventos, ou seja, a relação entre um

evento X (causa) e um evento Y (efeito) [Radinsky et al. 2012].

Mais recentemente, a análise de eventos tem ganhado destaque devido à popularização de plataformas web para publicação de conteúdo, especialmente portais de notícias, redes sociais, blogs e fóruns [Hou and Li 2015]. Essas plataformas armazenam eventos por meio de textos a respeito de diversos setores da sociedade e podem ser vistas como uma representação digital (mundo virtual) dos eventos que ocorrem em nosso mundo real [Radinsky and Horvitz 2013]. Nesse sentido, pesquisas computacionais envolvendo mineração de textos e aprendizado de máquina são importantes para extrair eventos a partir dos textos e então mapear os eventos desse mundo virtual para nosso mundo físico, o que permite a realização de estudos práticos envolvendo a análise de grandes bases de eventos [Hogenboom et al. 2016].

Dentre diversos métodos para apoiar a análise de eventos, métodos de agrupamento são uma estratégia interessante para organizar eventos em grupos, de forma que eventos alocados em um mesmo grupo sejam relacionados entre si [Conrad and Bender 2016; Florence et al. 2017]. Além disso, o agrupamento pode ser realizado de forma hierárquica, obtendo-se grupos e subgrupos de eventos. Dessa forma, a análise de eventos pode ser realizada em diversos níveis de granularidade. Outra importante característica é que o agrupamento de eventos representa um tipo de aprendizado não supervisionado, ou seja, exige pouco esforço humano para o aprendizado de um modelo de agrupamento entre os eventos [Aggarwal 2018].

Um dos principais desafios em agrupamento de eventos é definir uma medida de proximidade que identifique adequadamente quando dois ou mais eventos são relacionados entre si [Allan 2012; Radinsky et al. 2012; Radinsky and Horvitz 2013; Conrad and Bender 2016; Florence et al. 2017]. A maioria dos trabalhos existentes exploram a similaridade textual entre os eventos, com o pressuposto de que eventos de conteúdo similar podem estar relacionados entre si. Além disso, as informações de tempo e local de ocorrência entre eventos são utilizadas para restringir o cálculo da similaridade para eventos publicados em um determinado período e região. Embora seja uma abordagem interessante, a definição dessas restrições não é uma tarefa fácil, pois exige que o usuário defina *a priori* informações sobre duração e locais de ocorrência de eventos relacionados, informação que geralmente não está disponível [Hou and Li 2015; Conrad and Bender 2016; Florence et al. 2017]. Por exemplo, muitos eventos apresentam o comportamento de propagação em cadeia, como epidemias, o que dificulta definir regiões de interesse. Um problema similar acontece quando há interesse em identificar sazonalidade na ocorrência de eventos, análise que pode ficar prejudicada com uso de restrições de tempo.

Nesse trabalho é apresentada uma abordagem para agrupamento hierárquico e multivisão de eventos extraídos de textos. Nesse caso, as diferentes informações sobre os eventos, como informação textual, informação temporal e informação geográfica são consideradas diferentes visões durante a tarefa de agrupamento. Ao contrário das abordagens existentes, na proposta aqui apresentada as restrições de tempo e local são aprendidas automaticamente explorando os padrões extraídos no conjunto de eventos. Para lidar com o agrupamento hierárquico e multivisão, é proposta e avaliada uma estrutura denominada grafo de consistência que representa o consenso de agrupamentos entre as diferentes visões. O grafo de consistência proposto também é uma forma de indicar as relações entre pares de eventos, sendo uma estrutura útil para apoiar a análise de causalidade entre eventos. Para analisar a eficácia da proposta, foi realizada uma avaliação experimental com oito conjuntos de eventos de *benchmark*. Os resultados revelaram que a abordagem proposta é superior à abordagem tradicionalmente utilizada na área, apresentando ainda o diferencial de permitir a visualização das relações entre eventos por meio do grafo de consistência.

2. TRABALHOS RELACIONADOS

Na área de mineração de textos e aprendizado de máquina, um evento é comumente definido como algo que ocorre em determinado tempo e local [Allan 2012]. Um dos trabalhos pioneiros para análise de

eventos extraídos a partir de textos foi publicado por [Yang et al. 1998], que foi um projeto financiado pela DARPA (*Defense Advanced Research Projects Agency*), com o objetivo de construir uma base de conhecimento com eventos relacionados a partir de diversas fontes de notícias da web. Nesse trabalho, um evento i é representado em um espaço m-dimensional $e_i = (t_1, t_2, ..., t_m)$, composto por m termos, em que t_j indica o peso do termo j para o evento i, como ausência ou presença, frequência ou uso de ponderação TFIDF (*Term Frequency - Inverse Document Frequency*) [Aggarwal 2018] cuja frequência do termo no documento é ponderada pelo inverso do número de documentos em que o termo ocorre.

Nos primeiros trabalhos envolvendo análise de eventos, os termos eram representados por palavras-chave extraídas do texto dos eventos. Em trabalhos posteriores, foram investigadas formas de diferenciação entre esses termos por meio de componentes dos eventos, como informação textual, informação temporal e informação geográfica [Horie et al. 2016]. Em especial, destacam-se representações que organizam os termos do evento em múltiplas visões, de forma que os termos possam indicar componentes do tipo *what* (o quê?), *when* (quando?) e *where* (onde?). Dessa forma, um evento $e_i = (t_1^{v(1)}, t_2^{v(1)}, ..., t_1^{v(2)}, t_2^{v(2)}, ..., t_1^{v(3)}, t_2^{v(3)}, ...)$ possui mais de um tipo de visão $v(l)$, possibilitando ser diferenciadas durante o cálculo da proximidade entre dois eventos.

Nos últimos anos foram propostas medidas de proximidade que consideram diferentes visões de um objeto, sendo útil para tarefas de agrupamento de eventos [Radinsky and Horvitz 2013; Deza 2014]. A Equação 1 define uma medida de similaridade entre dois eventos e_i e e_j baseada em três componentes (*what, when* e *where*), na qual os termos α, β e γ indicam a importância de cada componente. Uma vez definida a medida de similaridade, então diversos algoritmos de agrupamento (particional e hierárquico) podem ser empregados.

$$sim(e_i, e_j) = \alpha sim^{v(what)}(e_i, e_j) + \beta sim^{v(when)}(e_i, e_j) + \gamma sim^{v(where)}(e_i, e_j) \quad (1)$$

Uma das principais críticas ao uso de uma única medida de proximidade para combinar as diferentes visões dos eventos é a dificuldade em lidar com a escala individual da medida de similaridade em cada visão, bem como definir seus respectivos níveis de importância. Além disso, a medida de similaridade de cada visão contém um conjunto de parâmetros para serem definidos conforme o domínio da aplicação, como limiar mínimo de similaridade de conteúdo, granularidade temporal e limiar mínimo para distância geográfica entre dois eventos. O conjunto de todos esses parâmetros torna o uso dessa abordagem muito difícil na prática, pois exige que usuários definam *a priori* informações sobre os eventos que, na prática, pretendiam obter *a posteriori* como resultado da própria análise da base de eventos.

Abordagens mais recentes exploram métodos de agrupamento que permitem a interação com o usuário para minimizar a complexidade de definição dos parâmetros de cada componente. No trabalho de [Conrad and Bender 2016], usuários fornecem exemplos de eventos de interesse que são utilizados como semente para obtenção do agrupamento de forma semissupervisionada. De forma similar, no trabalho de [Florence et al. 2017], usuários fornecem restrições de granularidade temporal e geográfica (raio máximo permitido com a distância entre dois eventos). Tais restrições são propagadas e incorporadas no algoritmo de agrupamento hierárquico por meio de restrições *cannot-link* (eventos que não devem ser agrupados) e *must-link* (eventos que devem ser agrupados) e, então, é utilizado um método de agrupamento semissupervisionado para organização dos eventos em grupos e subgrupos. Embora as abordagens semissupervisionadas reduzam a complexidade na definição dos parâmetros das componentes, é interessante que tal etapa possa ser realizada de forma automática a partir dos padrões existentes na própria base de eventos.

Neste trabalho, há interesse em investigar métodos de aprendizado multivisão como uma alternativa para lidar com as múltiplas componentes de um evento. Em vez de combinar diversas medidas de similaridades em uma única medida, o aprendizado multivisão visa obter uma hipótese (e.g. um modelo de agrupamento) em cada visão dos dados e então maximizar a concordância entre hipóteses

distintas [Zhao et al. 2017]. Embora existam diversas abordagens para agrupamento multivisão, a grande maioria é proposta apenas para agrupamento particional. Além disso, não foi encontrado na literatura estratégias de agrupamento hierárquico e multivisão para lidar com as especificidades do domínio de eventos.

A motivação desta proposta é que ao considerar um agrupamento individual em cada visão será possível identificar padrões específicos para as componentes *what*, *when* e *where*. Em seguida, a maximização do consenso entre cada modelo de agrupamento permite considerar tais padrões no modelo de agrupamento final, eliminando a necessidade de o usuário definir os parâmetros das componentes. Para tal, é proposta neste trabalho a estrutura de um grafo de consistência entre eventos como uma forma de maximizar a concordância entre diferentes componentes.

3. ABORDAGEM PROPOSTA

3.1 Formulação do Problema

O resultado de um modelo de agrupamento hierárquico $\mathbf{H}$ para uma base com n eventos pode ser representado por meio de uma matriz

$$\mathbf{H} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = a_{ij} \in \mathbb{R}^{n \times n} \tag{2}$$

em que a_{ij} indica o nível em que os eventos e_i e e_j foram agrupados na hierarquia. Dois eventos não similares só serão unidos no grupo raiz (que contém todos os eventos) e assim receberá o valor $a_{ij} = 0$.

Dessa forma, considere que $\mathbf{H}^{v(what)}$, $\mathbf{H}^{v(when)}$ e $\mathbf{H}^{v(where)}$ sejam os modelos de agrupamento hierárquico em cada componente da base de eventos, ou seja, em cada visão. Um agrupamento multivisão $\mathbf{H}^*$ é de tal forma que minimiza a discordância entre os agrupamentos hierárquicos individuais conforme definido na Equação 3, em que $C = \{what, when, where\}$ são as componentes do evento, $a_{ij}^{v(c)}$ indica o nível de agrupamento entre os eventos e_i e e_j na componente $c \in C$, e a_{ij}^* indica o nível de agrupamento entre os eventos e_i e e_j no agrupamento hierárquico multivisão.

$$\rho(\mathbf{H}^*) = \sum_{c \in C} \sum_{i,j=1}^{n} \| a_{ij}^{v(c)} - a_{ij}^* \|^2 \tag{3}$$

Encontrar a matriz $\mathbf{H}^*$ que minimiza a função $p(\mathbf{H}^*)$ é um problema computacionalmente difícil (*NP-Hard*), sendo necessário o uso de abordagens iterativas e/ou heurísticas para obter soluções em tempo aceitável, porém convergindo a um ótimo local. Nas próximas seções são descritas as etapas da abordagem aqui proposta, tanto para agrupamento hierárquico em cada visão quanto para a etapa de aprendizado multivisão via grafo de consistência.

3.2 Agrupamento nas Visões Textual (what), Temporal (when) e Geográfica (where)

A primeira etapa da abordagem proposta consiste em obter um modelo de agrupamento para cada visão do conjunto de eventos. Na prática, isso significa definir uma medida de similaridade para cada componente e, em seguida, aplicar algum método de agrupamento. Para a visão textual (componente *what*) é utilizada a medida de similaridade cosseno, definida na Equação 4, que explora as palavras-chave do texto de dois eventos e_i e e_j.

$$sim^{what}(e_i, e_j) = \frac{e_i \cdot e_j}{\|e_i\|\|e_j\|} \tag{4}$$

Para a visão temporal, o objetivo é agrupar eventos que ocorreram em períodos próximos, geralmente com base na data de publicação. No entanto, durante o pré-processamento algumas expressões temporais podem ser normatizadas para extração de datas, indicando que um evento pode ter uma ou mais datas associadas. A similaridade é baseada na diferença de marca temporal *timestamp*, definida na ISO 8601[1], conforme a Equação 5 que utiliza as datas de dois eventos e_i e e_j. A função $TS(e_i)$ retorna um conjunto de *timestamps* de um evento e_i. Observe que, neste caso, se um evento possui duas ou mais datas associadas, então a distância temporal é dada pela distância temporal média.

$$sim^{when}(e_i, e_j) = \frac{1}{|TS(e_i)|.|TS(e_j)|} \sum_{q \in TS(e_i)} \sum_{r \in TS(e_j)} \frac{1}{\|q - r\|^2 + 1} \tag{5}$$

Em relação à visão geográfica, o objetivo é agrupar eventos que ocorreram em regiões próximas conforme coordenadas de latitude e longitude. Tais coordenadas são extraídas do texto dos eventos por meio de pré-processamento para reconhecimento de entidades nomeadas (nomes de locais) seguido de um processo de *geocoding*, que consulta uma base de dados expressões georreferenciada[2]. A similaridade entre dois eventos utiliza como base a métrica de Haversine, que estima a distância geográfica entre duas coordenadas em uma esfera. A métrica de Haversine é convertida em similaridade na Equação 6, que se aproxima de 1 quando duas coordenadas estão próximas e de 0, caso contrário. As funções sin, arcsin e cos são funções trigonométricas. Já as variáveis lat_{e_i} e lon_{e_j} representam as coordenadas latitude e longitude, respectivamente, de um evento e_i (e de forma análoga para e_j).

$$sim^{where}(e_i, e_j) = \frac{1}{1 + 2r \arcsin\left(\sqrt{\sin^2\left(\frac{lat_{e_i} - lat_{e_j}}{2}\right) + \cos(lat_{e_i})\cos(lat_{e_j})\sin^2\left(\frac{lon_{e_i} - lon_{e_j}}{2}\right)}\right)} \tag{6}$$

Ao definir medidas de similaridade apropriadas para cada visão, os agrupamentos $\mathbf{H}^{v(what)}$, $\mathbf{H}^{v(when)}$ e $\mathbf{H}^{v(where)}$ podem ser obtidos considerando padrões da base de eventos sem a necessidade de definição de parâmetros de limiares para cada componente. Na próxima seção, os agrupamentos individuais são combinados por meio de aprendizado multivisão.

3.3 Aprendizado Multivisão via Grafo de Consistência

Neste trabalho é proposta uma estrutura denominada grafo de consistência para obter uma solução aproximada do problema de agrupamento multivisão. Considere um grafo $G = (V, E, \mathbf{W})$, em que $V = \{e_1, e_2, ..., e_n\}$ indica um conjunto não vazio de eventos (vértices), $E = \{(e_i, e_j)\}\ \forall i \neq j$ indica um conjunto de relação de pares de eventos (arestas) e $\mathbf{W}$ é uma matriz que indica o peso de cada aresta (relação entre eventos).

O objetivo do grafo de consistência é identificar pares de eventos que foram alocados nos mesmos grupos em todos os modelos de agrupamento, ou seja, nas componentes *what*, *when* e *where*. A ideia básica é que se dois eventos possuem conteúdo similar, foram publicados em mesmo intervalo temporal e ocorreram em regiões próximas, então há uma maior probabilidade de que estejam relacionados. O peso da aresta indica, então, a consistência da relação entre dois eventos.

[1]ISSO 8601 é um padrão internacional de definição de marcação temporal emitida pela Organização Internacional de Padronização.
[2]Nesse trabalho foi utilizada a API de Geocoding do projeto Websensors: `https://websensors.net.br/`

O aprendizado do grafo de consistência é realizado pela construção da matriz de pesos $\mathbf{W}_{n\times n}$ para cada par de arestas, conforme Equação 7, que utiliza como entrada os modelos de agrupamento hierárquico $\mathbf{H}^{v(what)}$, $\mathbf{H}^{v(when)}$ e $\mathbf{H}^{v(where)}$ de cada visão, em que $C = \{what, when, where\}$.

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} = w_{ij} \in \mathbb{R}^{n\times n},\ w_{ij} = \sum_{c\in C}\sum_{i,j=1}^{n} consistency(a_{ij}^{v(c)}) \tag{7}$$

A função *consistency*(.) da Equação 7 tem a finalidade de verificar se os eventos e_i e e_j foram alocados nos mesmos grupos em todas as visões, respeitando-se um determinado nível da hierarquia. A literatura dedicada à análise de agrupamento discute que o nível de agrupamento pode ser superestimado pelo piso $\lfloor\sqrt{n}\rfloor$ (menor número inteiro menor ou igual a $\sqrt{n}$) em que n é o número de eventos. Por exemplo, ao considerar $n = 100$ eventos, dois eventos serão conectados no grafo de consistência se, e somente se, tais eventos foram alocados juntos nas três visões *what*, *when* e *where* a partir do nível 10 em todas as hierarquias. Caso contrário, a função *consistency*(.) irá retornar o valor 0 para indicar que não há relação entre esse par de eventos.

O agrupamento hierárquico e multivisão final $\mathbf{H}^*$, que minimiza a discordância entre as múltiplas visões dos eventos, é obtido aplicando-se um método de agrupamento hierárquico a partir da matriz de pesos $\mathbf{W}$. É importante observar que o valor 0 indica ausência de aresta no grafo de consistência. Assim, na etapa de agrupamento hierárquico, tais relações devem ser desconsideradas, o que pode gerar estruturas de árvores não binárias. Outra observação importante é que, na prática, $\mathbf{W}$ pode ser vista como uma matriz de similaridades entre pares de eventos que foram agrupados juntos, de forma consistente, em todas as visões. Assim, é uma solução aproximada para o problema formulado na Equação 3.

4. AVALIAÇÃO EXPERIMENTAL

Para avaliar a eficácia da abordagem proposta, foi conduzida uma avaliação experimental envolvendo oito conjuntos de dados de eventos de *benchmark* proveniente da Reuters RCV2[3]. Nesses conjuntos de dados, cada evento possui uma categoria manualmente rotulada pela Reuters. Além disso, os eventos possuem informação sobre data de publicação, palavras-chave e informação geográfica. Um sumário dos conjuntos de dados utilizados é apresentado na Tabela I.

Table I. Visão geral dos conjuntos de dados utilizados para a avaliação experimental.

Conjunto de dados	**#Eventos**	**#Atributos**			**#Categorias**
		What	*When*	*Where*	
BUSINESS TRANSACTIONS (BT)	17802	409	362	189	4
CONSUMER FINANCES (CF)	1085	79	276	66	3
INFLATION (INF)	2126	46	321	150	2
INVESTMENTS (INV)	19064	206	349	150	4
LAWSUITS (LAW)	19543	568	365	208	2
NATURAL DISASTERS (ND)	12582	411	364	231	3
REPORTS (REP)	22079	563	365	227	4
TRADE RESERVES (TR)	8850	209	361	217	3

Dado um agrupamento hierárquico $\mathbf{H}$, a qualidade da solução é calculada por meio de um critério de acurácia conforme definido da Equação 8. Nessa medida, $\#ParesCorretamenteAgrupados$ indica

[3]Reuters RCV2: `https://trec.nist.gov/data/reuters/reuters.html`

a quantidade de pares de eventos que foram corretamente agrupados (e.g. da mesma categoria) e $\#TotalParesReferencia$ indica o número máximo de pares de eventos que deveriam ser agrupados considerando a informação manualmente rotulada do conjunto de dados — desconsiderando possíveis repetições e a raiz da hierarquia.

$$ACC = \frac{\#ParesCorretamenteAgrupados}{\#TotalParesReferencia} \tag{8}$$

A abordagem proposta foi comparada com um método tradicional na área que utiliza uma única medida de similaridade para combinar todas as componentes (conforme discutido na Seção 2). Nesse caso, a definição dos parâmetros para as componentes de cada visão foi baseada em análies experimental de trabalhos anteriores [Conrad and Bender 2016; Florence et al. 2017]: um limiar de 15 dias para similaridade temporal e um limiar de no máximo 500km para similaridade geográfica entre dois eventos. Foi definido o mesmo peso de importância para todas as visões. Em todos os experimentos foi utilizado o método de agrupamento hierárquico UPGMA, considerado um dos estado-da-arte para dados textuais.

Na Tabela II é apresentada uma comparação experimental entre a abordagem proposta e a abordagem de referência. Dentre 8 conjuntos de dados de eventos analisados, a abordagem proposta apresentou maior acurácia em 6 conjuntos de dados. O único conjunto de dados em que a abordagem proposta obteve acurácia inferior foi o *Inflation (INF)* que, por apresentar um pequeno conjunto de eventos, não foi eficaz na construção do grafo de consistência. Por outro lado, em geral há um significativo percentual de melhora em relação à abordagem de referência, chegando em até 29% de melhora para o conjunto de dados *Reports (REP)*.

Table II. Comparação experimental entre a abordagem proposta e a abordagem de referência da literatura.

Abordagem	**BT**	**CF**	**INF**	**INV**	**LAW**	**ND**	**REP**	**TR**
Referência	0.93	0.90	0.98	0.91	0.90	0.87	0.62	0.93
Proposta	0.95	0.96	0.88	0.97	0.90	0.93	0.80	0.98
% de Melhora	2.2	6.7	-10.2	6.6	0.0	6.9	29.0	5.4

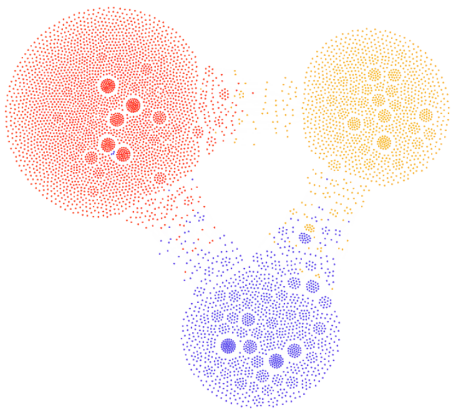

Fig. 1. Visualização do grafo de consistência para o conjunto de dados *Natural Disasters (ND)*. Cada evento foi colorido de forma a distinguir suas diferenste categorias.

A melhoria obtida com a abordagem proposta é proveniente principalmente do grafo de consistência entre eventos, que tem a vantagem de se adaptar conforme os padrões de cada visão, ao invés de depender dos parâmetros (limiares) a serem informados antes do processo. Tal estrutura também

define de forma implícita a importância de cada visão. Quando uma determinada visão não apresenta uma estrutura de grupo adequada para um subconjunto de eventos, tais eventos naturalmente são eliminados do grafo de consistência. Para exemplificar a qualidade do grafo de consistência, na Figura 1 é ilustrada a relação entre pares de eventos considerados consistentes para o conjunto de dados *Natural Disasters (ND)*[4]. É possível observar que a ideia de relacionar no grafo de consistência apenas os eventos que concordam em agrupamentos provenientes das três componentes *what*, *when* e *where* é uma estratégia robusta, ao menos, para identificar eventos (nós) de uma mesma categoria (indicada por meio de cores distintas na visualização).

5. CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentada uma nova abordagem para agrupamento hierárquico e multivisão como uma alternativa para análise exploratória de eventos. Dentre as principais contribuições, vale destacar a proposta de um grafo de consistência como estrutura para combinar modelos de agrupamento provenientes das múltiplas visões dos eventos, como informação textual, temporal e geográfica.

Os resultados experimentais indicam que a abordagem proposta é promissora, obtendo acurácia de agrupamento superior à abordagem tradicional utilizada. Além disso, a presente proposta possui a vantagem de realizar agrupamento de eventos sem a necessidade de definir parâmetros relacionados aos limiares de similaridade temporal e geográfico, que geralmente é uma tarefa árdua para os usuários.

As direções para trabalhos futuros envolvem incorporar outras visões dos eventos, como nomes de pessoas, organizações e tópicos latentes. Ainda, espera-se avaliar o impacto de diferentes métodos de agrupamento hierárquico na análise de eventos.

6. AGRADECIMENTOS

Este trabalho contou com o apoio das seguintes agências de fomento: FAPESP (Processo 2017/08804-2), Fundect-MS (Processo 14/08996-0), CAPES, CNPq e FINEP. Os autores agradecem a NVIDIA pela doação de GPUs (*GPU Grant Program*).

REFERENCES

Aggarwal, C. C. *Machine learning for text*. Springer, 2018.

Allan, J. *Topic detection and tracking: event-based information organization*. Vol. 12. Springer, 2012.

Conrad, J. G. and Bender, M. Semi-supervised events clustering in news retrieval. In *Recent Trends in News Information Retrieval Workshop*. pp. 21–26, 2016.

Deza, M. M. Distances and similarities in data analysis. In *Encyclopedia of Distances*. Springer, pp. 323–339, 2014.

Florence, R., Nogueira, B., and Marcacini, R. Constrained hierarchical clustering for news events. In *Proceedings of the 21st International Database Engineering & Applications Symposium*. ACM, pp. 49–56, 2017.

Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F., and Caron, E. A survey of event extraction methods from text for decision support systems. *Decision Support Systems* vol. 85, pp. 12–22, 2016.

Horie, S., Kiritoshi, K., and Ma, Q. Abstract-concrete relationship analysis of news events based on a 5W representation model. In *Int. Conference on Database and Expert Systems Applications*. Springer, pp. 102–117, 2016.

Hou, L. and Li. Newsminer: multifaceted news analysis for event search. *KBS Journal* vol. 76, pp. 17–29, 2015.

Radinsky, K., Davidovich, S., and Markovitch, S. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, pp. 909–918, 2012.

Radinsky, K. and Horvitz, E. Mining the web to predict future events. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. ACM, pp. 255–264, 2013.

Yang, Y., Pierce, T., and Carbonell, J. A study of retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference*. ACM, pp. 28–36, 1998.

Zhao, J., Xie, X., Xu, X., and Sun, S. Multi-view learning overview: Recent progress and new challenges. *Information Fusion* vol. 38, pp. 43–54, 2017.

[4]A visualização foi obtida por meio da biblioteca VIS.js (`http://visjs.org/`)