Integração de Dados Factuais e Subjetivos: Um Estudo de Caso em Comércio Eletrônico

Altigran Soares da Silva[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Av. General Rodrigo Octavio Jordão Ramos, 1200 – Manaus – Brasil

Resumo. *Ao longo das últimas décadas os bancos de dados (BDs) têm sido o principal recurso computacional utilizado para armazenamento e gerenciar dados dos mais variados tipos de aplicações. Tipicamente, BDs armazenam informações factuais e objetivas sobre entidades do mundo real, que são representadas como um conjunto de atributos. No entanto, tem havido recentemente uma grande demanda para representar em BDs conhecimento subjetivo sobre estas entidades, com objetivo final de permitir que usuários e aplicações possam lidar com informações subjetivas de forma mais eficaz. Neste artigo, discutimos o problema de integração de dados factuais e subjetivos em bancos de dados, utilizando para isso um estudo de caso em comércio eletrônico. Neste domínio, a integração de informações subjetivas fornecidas por usuários nas mídias sociais com dados factuais fornecidos por vendedores e fabricantes de produtos tem o potencial de gerar conhecimento relevante para uma variedade de aplicações. Especificamente, exploramos neste artigo essa ideia em um estudo realizado para entender os interesses dos consumidores de produtos eletrônicos em sites de comércio eletrônico.*

Abstract. *Over the last decades, databases (DBs) have been the main computational resource used for storing and managing data in a variety of application types. Typically, DBs store factual and objective information about real-world entities, which are represented as a set of attributes. However, there has recently been a great demand to also represent subjective knowledge about these entities in DBs, with the ultimate goal of allowing users and applications to deal with subjective information more effectively. In this article, we discuss the problem of integrating factual and subjective data into databases, using a case study on electronic commerce. In this domain, the integration of subjective information provided by users on social media with factual data provided by vendors and product manufacturers has the potential to generate relevant knowledge for many downstream applications. Specifically, in this article we explored this idea in a study carried out to understand the interests of consumers of electronic products on e-commerce sites.*

1. Introdução

Nas últimas décadas os bancos de dados têm sido o principal recurso computacional utilizado para armazenamento e gerencia de dados nos mais variados tipos de organizações. Tipicamente, BDs armazenam informações factuais e objetivas sobre entidades do mundo real. Por exemplo, um BD que implementa um catálogo de um site de comércio eletrônico armazena informações objetivas (factuais), tais como o tamanho da tela de

um smartphone, sua quantidade de memória, o modelo de seu processador, etc. O objetivo é informar aos clientes as características do produto, que são representadas como um conjunto de atributos previamente definidos no banco de dados. Estas informações constituem fontes valiosas para apoiar a tomada de decisão de compra. De acordo com Park et al. [Park et al. 2012], os atributos do produto informados nos sites encorajam o comportamento de navegação do consumidor, o que muitas vezes pode levar à compra por impulso. Portanto, os atributos do produto são um elemento crucial que influencia a escolha do produto do cliente [Kostyra et al. 2016].

Por outro lado, usuários em geral também têm demonstrado interesse em consultar informações subjetivas sobre as entidades representadas no BD. Por exemplo, os usuários podem estar interessados na opinião de outros usuários sobre o desempenho do processador do smartphone, se o tamanho é considerado adequado para mulheres, etc. A importância de considerar informações subjetivas além das informações factuais já foi verificada em muitas aplicações relacionadas ao comércio eletrônico. De fato, é uma prática comum considerar as opiniões de outras pessoas antes de comprar um determinado produto, já que existem milhares de resenhas feitas por usuários disponíveis online. Segundo um levantamento recente[1], 82% dos americanos recorrem a resenhas online antes de comprar um produto pela primeira vez e 40% sempre recorrem a resenhas antes de boas compras. Outra pesquisa[2] com consumidores online que abrangeu 5 continentes diferentes, revela que 45% dos consumidores consideram que as resenhas são o recurso mais influente das mídias sociais no seu comportamento de compras. Essa pesquisa também revela que ler resenhas sobre produtos e empresas é a quarta atividade mais comum dos compradores dentro de lojas físicas. Na maioria dos casos, essas informações subjetivas são fornecidas por opiniões emitidas por outros clientes em resenhas. Essa tendência foi potencializada com surgimento da chamada Web 2.0, que permitiu o surgimento de uma grande quantidade de informações subjetivas (opinativas) disponíveis sobre os produtos e suas características. De fato, as mídias sociais on-line tornaram-se uma parte essencial de nossa vida diária. Por meio dessas mídias, os usuários geram e trocam informações usando diversos mecanismos de comunicação [Kaplan and Haenlein 2010]. Neste contexto, cada vez mais usuários difundem e confiam em opiniões publicadas por outros usuários sobre os mais diversos tópicos, incluindo aí opiniões e informações sobre produtos [Pang and Lee 2007].

Assim, no domínio de comércio, a integração de informações subjetivas implicitamente fornecidas por usuários nas mídias sociais com dados factuais fornecidas por vendedores e fabricantes de produtos tem o potencial de gerar conhecimento relevante para uma variedade de aplicações. Neste artigo, exploramos essa ideia em um estudo realizado para entender os interesses dos consumidores de produtos eletrônicos em sites de comércio eletrônico.

No entanto, o problema de integrar informações subjetivas com dados factuais é bem mais desafiante que as operações de junção típicas de bancos de dados tradicionais, ou mesmo que tarefas clássicas de integração de dados, como resolução de entidades e

[1]PricewaterhouseCoopers (2016). They say they want a revolution https://www.pwc.com/gx/en/retail-consumer/publications/assets/total-retail-global-report.pdf

[2]Smith, A.; Anderson, M. (2016). Online shopping and e-commerce. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2016/12/PI_2016.12.19_Online-Shopping_FINAL.pdf

casamento de esquemas [Doan et al. 2012]. Na Figura 1[3] ilustramos um exemplo deste tipo de integração. Na parte superior da figura, vemos um conjunto de sentenças que compõem resenhas escritas por usuários. Cada uma destas sentenças pode ter uma ou mais opiniões, que são identificadas na figura com números. O processo de integração propriamente dito é ilustrado na parte central da figura, onde as opiniões são mapeadas para os atributos de produto aos quais elas se referem. Note-se que o mapeamento não é óbvio, pois o atributo referenciado não é mencionado explicitamente nas opiniões. Na parte inferior da figura, ilustramos uma tabela resultante do processo de integração, onde a representação dos produtos inclui, além os dados factuais usuais, uma nova dimensão que adiciona informações subjetivas fornecidas pelas opiniões dos usuários.

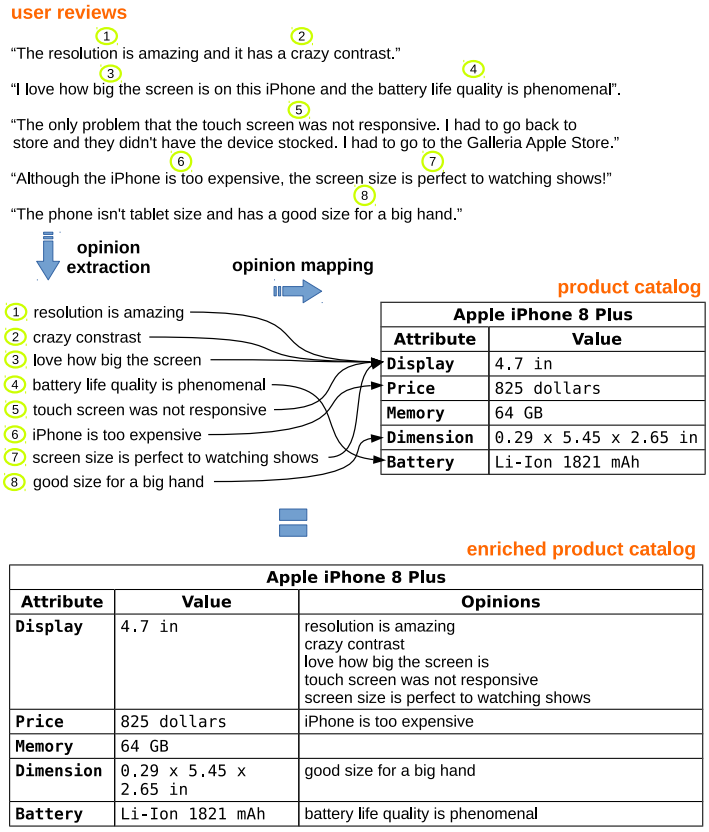

Figura 1. Exemplo de integração de informações subjetivas com dados factuais.

Esse nova representação pode dar origem a uma versão enriquecida do catálogo de produtos. Além disso, ela permite também estruturar informações e conhecimento de alta relevância para uma variedade de aplicações de grande interesse como: (1) a previsão de comportamento de grupos de usuários; (2) o fornecimento de recomendações confiáveis e de alta qualidade; (3) estabelecimento de preços que maximizem interesses de fornecedores e consumidores; (4) a apresentação sintetizada de opiniões de comunidades e (5) busca de comentários relacionado com um certo conteúdo, produto ou serviço. Além disso, o fato destes conteúdo estarem publicamente disponíveis torna possível a estimativa automática, entre outras coisas, das opiniões veiculadas, da polaridade da visão dos usuários (negativa, positiva ou neutra) e das suas emoções em relação às diferentes

[3]Dados extraídos do site da Amazon.com em língua inglesa.

entidades que são mencionadas ou aos aspectos destes produtos.

Neste artigo, discutimos o problema de integração de dados factuais e subjetivos em bancos de dados, utilizando para isso um estudo de caso em comércio eletrônico. Na Seção 2, descrevemos este estudo de caso. Na Seção 3 apresentamos a metodologia o usada. Na Seção 4, apresentamos e discutimos os resultados do estudo. Finalmente, na Seção 5, descrevemos trabalhos relacionados.

2. Integração de Dados Factuais e Subjetivos: Um Estudo de Caso

Como um exemplo do potencial da integração de dados factuais e subjetivos, nesta seção apresentamos um estudo cujo objetivo é investigar o impacto dos atributos, que constituem dados factuais sobre produtos, nas opiniões dos usuários, que constituem informações subjetivas. Utilizamos para isso resenhas disponíveis online escritas por usuários sobre produtos eletrônicos de sites de comércio eletrônico.

Uma resenha é um texto postado por um usuário em um site de comércio eletrônico, geralmente relatando sua experiência com um produto específico, que chamamos de *produto-alvo* da resenha. Cada resenha é composta por um conjunto de sentenças. As sentenças que expressam informações factuais são chamadas de sentenças *objetivas*, enquanto as sentenças que expressam crenças ou sentimentos pessoais são chamadas de sentenças *subjetivas* ou *opinativas* [Liu 2015]. Estamos interessados nas sentenças opinativas que representam as opiniões do usuário sobre um produto. Uma sentença opinativa pode ainda ser classificada como *comparativa* ou *direta*. Uma sentença comparativa expressa uma opinião sobre semelhanças ou diferenças entre dois ou mais produtos. A sentença "a câmera do iPhone é muito melhor que o Galaxy" é um exemplo de sentença comparativa. Uma *sentença opinativa direta* ou *direct opinionated sentence* (DOS) expressa uma opinião direta sobre uma característica ou parte do produto, ou sobre o produto como um todo. A frase "a câmera do iPhone é fantástica" é um exemplo de DOS. Como nosso objetivo é investigar o impacto dos atributos do produto em relação ao produto-alvo, consideramos apenas sentenças opinativas diretas, uma vez que as opiniões em sentenças comparativas são relativas. Para análises de usuários online postadas em sites de comércio eletrônico, esperamos obter um grande volume de DOS.

Ao analisar comentários reais de usuários, percebemos que os comentários também incluem opiniões que não se referem a um atributo específico do produto, mas ao produto como um todo. Além disso, as opiniões também podem referir-se a atributos que não estão representados no catálogo de produtos. Assim, consideramos que as opiniões podem ter três alvos distintos: a) `Atributo`, quando a opinião é sobre um atributo específico do produto; b) `Geral`, quando a opinião for sobre o produto como um todo; e, c) `Outro`, quando a opinião for sobre uma característica do produto alvo que não se apresenta como atributo. Apesar disso, esperamos que os atributos tenham um efeito persuasivo nas resenhas dos usuários.

3. Metodologia

Para o nosso estudo, construímos um conjunto de dados experimental a partir de uma grande coleção de cerca de 142 milhões de resenhas previamente coletadas do site Amazon.com, e selecionamos todas as resenhas de produtos em cinco categorias: câmeras (CAM), telefones celulares (CEL), DVD players (DVD), laptops (LAP) e roteadores de

Internet (ROT). Cada resenha nesta coleção identifica o produto ao qual elas se refere. A Tabela 1 apresenta, para cada categoria, o número de resenhas e sentenças dessa coleção, juntamente com a quantidade de produtos referenciados nas resenhas. Para completar nosso conjunto de dados experimental, coletamos da Amazon.com um conjunto de atributos usados em produtos de cada uma das cinco categorias. Esses conjuntos são mostrados na Tabela 2.

Categoria	Nr. Produtos	Nr. Resenhas	Nr. Sentenças
CAM	8.893	203.836	1.012.077
CEL	7.693	182.491	707.407
DVD	2.503	61.836	243.939
LAP	9.491	115.138	580.955
ROT	1.592	84.059	329.305
Total	**30.172**	**647.360**	**2.864.683**

Tabela 1. Dados de resenhas coletadas da Amazon.com

Categoria	Atributos de Produtos
CAM	*Dimension*, *Exposure Control*, *Imaging*, *Memory*, *Performance*, *Power*, *Price*, *Zoom*
CEL	*Battery*, *Camera*, *Dimension*, *Display*, *Memory*, *Price*, *Processor*, *Software*
DVD	*Accessory*, *Audio*, *Dimension*, *Price*, *Sound*, *Video*
LAP	*Battery*, *Connectivity*, *Dimension*, *Graphic*, *Memory*, *Price*, *Processor*, *Screen*, *Software*
ROT	*Accessory*, *Coverage Area*, *Dimension*, *Ports*, *Price*, *Security*, *Software*, *Speed*

Tabela 2. Atributos de produtos em cada categoria.

Como já foi dito, nosso estudo utiliza apenas sentenças opinativas diretas (DOS). Assim, desenvolvemos um método denominado *filterDOS* para selecionar DOSs nas resenhas. O filterDOS tem três etapas e permite identificar três tipos de sentenças: factual, comparativa e DOS. Na primeira etapa, as avaliações são divididas em sentenças. Na segunda etapa, o método identifica sentenças subjetivas obtidas na primeira etapa utilizando o método proposto por Qadir [Qadir 2009]. Por fim, o filterDOS elimina as sentenças comparativas, baseado no método proposto por Liu [Liu 2010].

A coleção de resenhas apresentadas na Tabela 1 tem mais de 2.800.000 sentenças. Portanto, mesmo após filtrar sentenças factuais e comparativas, o número de sentenças a serem analisadas ainda é enorme, com mais de um milhão de sentenças. Um conceito central na área de mineração de opiniões que usamos também em nosso estudo é o de *expressões de aspecto*. Consideramos que toda opinião é representada por uma expressão de aspecto [Liu 2015], onde um *aspecto* é qualquer referência feita em uma opinião a uma determinada parte ou característica do produto, ou mesmo ao produto como um todo. Como seria inviável anotar manualmente todos os aspectos encontrados neste enorme volume de sentenças, criamos um conjunto de dados experimental focado nas 100 expressões de aspecto mais frequentes encontradas nas DOSs das resenhas de cada categoria de produto. Argumentamos que utilizar algumas das expressões de aspecto mais frequentes é mais representativo do que considerar cada expressão de aspecto possível. Para selecionar as 100 expressões de aspecto mais frequentes, primeiro executamos o método de extração de aspecto proposto por Poria et al. [Poria et al. 2014]. Assim, todas as expressões de aspecto possíveis identificadas por esse método foram extraídas. Em seguida, classificamos essas expressões de acordo com sua frequência. Para garantir que usamos

apenas expressões de aspecto verdadeiras, inspecionamos manualmente as expressões extraídas usando a ordem de classificação e removemos aquelas que não consideramos como expressões de aspecto. No final, apenas as 100 verdadeiras expressões de aspecto mais frequentes foram mantidas para cada categoria de produto. Por fim, para cada categoria de produto, examinamos manualmente cada um dos 100 principais aspectos selecionados anteriormente e anotamos cada um com os atributos de produto mais relacionados a ele. Também anotamos os casos em que a opinião é sobre o produto como um todo (`Geral`) e quando a opinião é sobre uma característica do produto alvo que não é representada como um atributo canônico (`Outro`). A Table 3 apresenta um resumo do conjunto de dados experimentais que geramos, que mostra, para cada categoria, o número total de DOSs (DOSs - Total) e o número de DOSs que incluem pelo menos um dos 100 principais aspectos (DOSs - top 100 Aspectos) Como pode ser observado, esses DOSs representam mais de 50% de todos os DOSs.

Categoria	DOSs - Total	DOSs - top 100 Aspectos
CAM	476.605	249.714
CEL	277.712	138.939
DVD	89.525	48.608
LAP	189.782	126.865
ROT	123.336	73.027
Total	**1.156.960**	**637.153**

Tabela 3. Resumo do conjunto de dados experimental.

4. Resultados

4.1. Distribuição de Sentenças entre os Tipos de Alvos

No nosso conjunto de dados experimental apresentado na Tabela 3, mais de 1.100.000 sentenças são DOSs. O fato de uma grande fração das sentenças ser DOS mostra que sites de comércio eletrônico, como Amazon.com, são de fato úteis como uma fonte valiosa de opiniões sobre produtos alvo. Além disso, a partir do conjunto de dados experimental, podemos observar que mais de 55% dos DOSs contêm pelo menos um dos 100 principais aspectos. Essa observação corrobora nossa suposição de que usar as expressões de aspecto mais frequentes é mais adequado do que usar toda as expressões de aspecto. A Figura 2 sumariza a distribuição de sentenças entre os três tipos de alvos: `Atributos`, `Geral` e `Outro`. O número de sentenças contendo pelo menos um dos principais 100 aspectos que formam as opiniões são anotados com um tipo de alvo. Note-se que uma única sentença pode conter mais de uma opinião, e cada opinião pode referir-se a um tipo diferente de alvo. Assim, a soma da porcentagem de todos os tipos de alvos pode ser maior que 100Novamente, observamos que a maioria dos DOSs inclui expressões de aspecto que se referem a atributos canônicos do produto, identificados como Atributo. Por exemplo, nas categorias CAM e LAP, eles respondem por mais da metade das sentenças. Além disso, grande parte das sentenças contém opiniões referentes ao alvo "Geral".

4.2. Distribuição de Expressões de Aspecto

A Figura 3 mostra a distribuição dos 100 principais aspectos entre os três tipos de alvos. Observe que, para todas as categorias, a fração das expressões de aspecto que representam um atributo específico é superior a 50%. Isso mostra que a maioria das opiniões

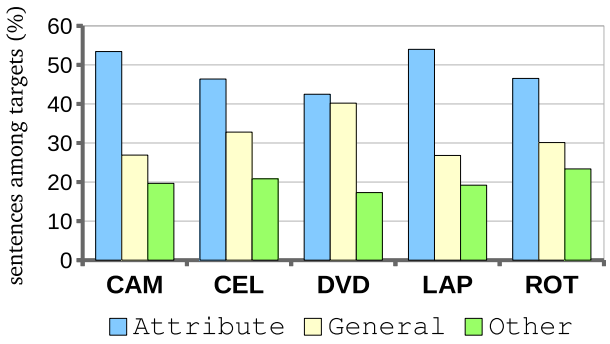

Figura 2. Distribuição das sentenças entre os tipos de alvo das opniÕes.

dos usuários postadas em sites de comércio eletrônico são sobre os atributos do produto. Além disso, pode-se observar que uma grande parte das expressões de aspecto refere-se ao alvo `Outro`. Por exemplo, nas categorias CAM, CEL e DVD, quase um quarto das expressões de aspecto estão em opiniões que foram anotadas como alvo `Outro`. Um problema intrigante que deixamos para trabalhos futuros é analisar mais a fundo esses casos de características latentes específicas que, embora não sejam representadas por algum atributo, são de interesse dos usuários. Por exemplo, "teclado" é a segunda expressão de aspecto mais frequente na categoria LAP, mas normalmente não há nenhum atributo referindo-se a ele nos atributos do produto. Em suma, a Figura 3 sugere que os usuários comentam com mais frequência sobre as características específicas dos produtos do que sobre o produto como um todo. Isso mostra a relevância de integrar adequadamente as opiniões subjetivas aos atributos factuais.

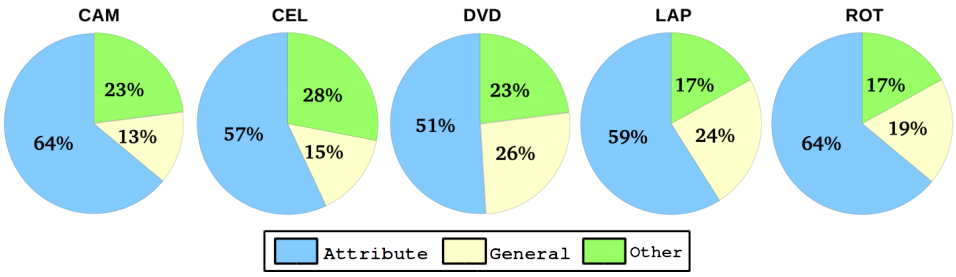

Figura 3. Distribuição dos 100 aspectos principais entre os três tipos de alvos.

4.3. Distribuição de Sentenças entre os Atributos

A Figura 4 apresenta a distribuição das sentenças entre os atributos de produtos para cada categoria. Nesses gráficos, cada vértice do polígono representa um atributo definido pelos fabricantes nas especificações do produto. O gráfico mostra a porcentagem de sentenças que contêm uma expressão de aspecto que corresponde a um determinado atributo. Em cada gráfico, os atributos são colocados no sentido horário, do mais para o menos referenciado. Por exemplo, mais de 40% das frases que incluem pelo menos um dos 100 principais aspectos da categoria DVD referem-se ao atributo "Acessório". Existem alguns atributos que são mencionados com muito mais frequência nas resenhas do que outros da mesma categoria. Por exemplo, na categoria CEL, os usuários comentam 2 vezes mais sobre a bateria do que sobre o preço dos telefones celulares. Esses resultados mostram

que existem certos atributos que são mais relevantes do que os outros em categorias específicas. Curiosamente, nas cinco categorias deste experimento, o preço não é o atributo mais comentado.

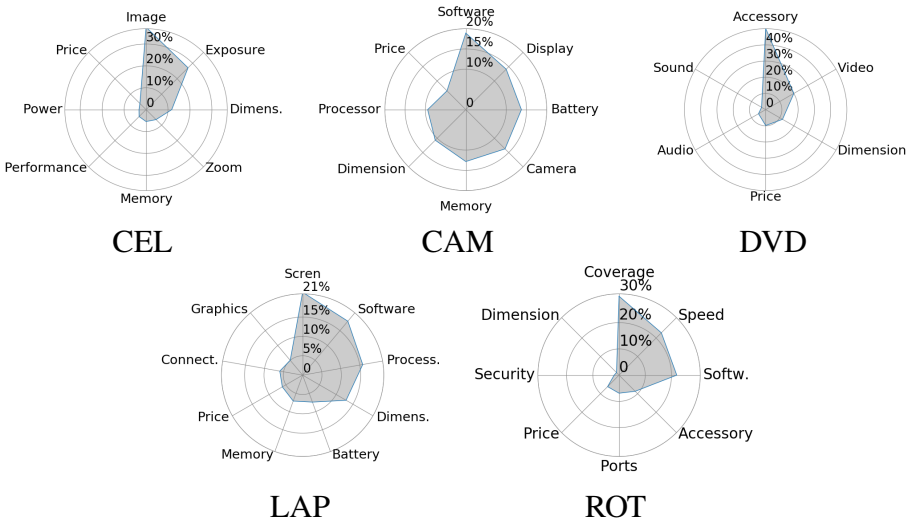

Figura 4. Distribuição das sentenças entre os atributos para cada categoria.

4.4. Diversidade de Expressões de Aspecto sobre Atributos

A Figura 5 mostra a distribuição das expressões de aspecto extraídas das resenhas dos usuários sobre os atributos em cada categoria. Nestes gráficos, mostramos o número de expressões de aspecto únicas que se referem ao mesmo atributo. Por exemplo, na categoria LAP, encontramos dez expressões de aspecto diferentes que se referem ao atributo ”Software”. Analisando as sentenças, descobrimos que os usuários realmente empregam vários termos diferentes, como “aplicativos”, “sistema”, “vista” e “programa” para se referir ao atributo ”Software”na categoria LAP. Esta avaliação experimental suporta mostra que os usuários costumam fazer menções diretas e indiretas aos atributos canônicos do produto. Para se ter uma ideia dos 100 principais aspectos mencionados, a Tabela 4 ilustra as dez expressões de aspecto mais frequentes extraídas nas avaliações de cada categoria, juntamente com o nome do atributo, quando se referem ao alvo ”Atributo”, ”Geral”ou ”Outro”. A partir desses resultados, fica aparente que as dez expressões de aspecto mais frequentes extraídas são bastante representativas de cada categoria de produto e, mais importante, os resultados mostram quais são os aspectos mais comentados relacionados aos atributos. Observe que a maioria das expressões de aspecto não correspondem exatamente ao nome do atributo do produto.

4.5. Observações

Neste estudo, verificamos que uma grande fração das sentenças em resenhas é composta por sentenças opinativas diretas, mostrando que os sites de comércio eletrônico são uma fonte valiosa de opiniões sobre os produtos. Utilizamos esse grande número de sentenças e, por meio de um protocolo bem definido, pudemos verificar que a maioria das opiniões dos usuários postadas em sites de comércio eletrônico se referem a um dos atributos do produto. Além disso, pudemos verificar que, em cada categoria, existem determinados atributos que são mais relevantes para os usuários do que os outros atributos. Por fim,

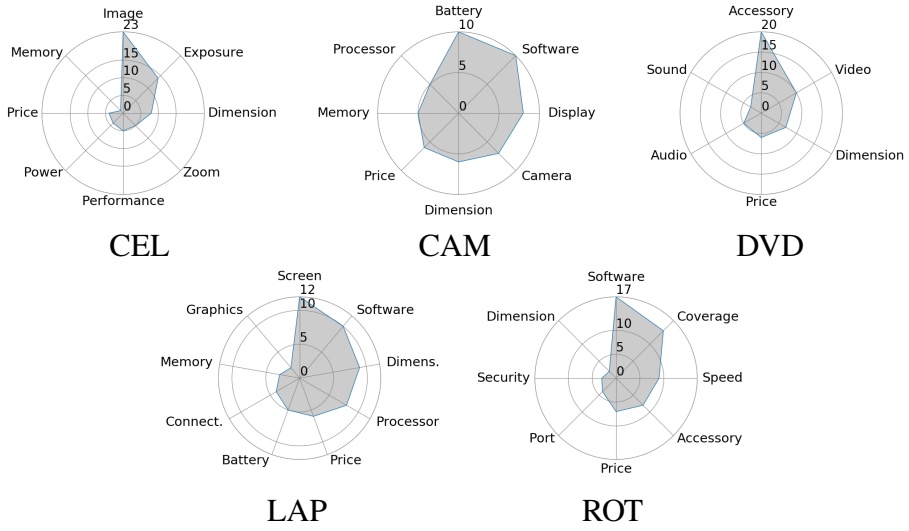

Figura 5. Distribuição das Expressões de Aspecto sobre os Atributos.

CAM	CEL	DVD	LAP	ROT
camera `(Geral)`	phone `(Geral)`	unit `(Geral)`	laptop `(Geral)`	router `(Geral)`
quality `(Geral)`	easy `(Outro)`	dvd `(Geral)`	keyboard `(Outro)`	easy `(Outro)`
picture *(Imaging)*	quality `(Geral)`	easy `(Outro)`	computer `(Geral)`	instructions `(Outro)`
lens *(Exposure Control)*	feature `(Outro)`	quality `(Geral)`	software *(Software)*	software *(Software)*
shots *(Exposure Control)*	card *(Memory)*	player `(Geral)`	fast *(Processor)*	unit `(Geral)`
features `(Other)`	size *(Dimension)*	picture quality *(Video)*	size *(Dimension)*	speed *(Speed)*
size *(Dimension)*	software *(Software)*	instructions `(Outro)`	card *(Memory)*	device `(Geral)`
card *(Memory)*	camera *(Camera)*	features `(Outro)`	screen *(Screen)*	internet *(Coverage Area)*
settings `(Geral)`	screen *(Display)*	picture *(Video)*	easy `(Outro)`	network *(Coverage Area)*
pics *(Imaging)*	keyboard `(Outro)`	product `(Outro)`	graphics *(Graphics)*	settings `(Outro)`

Tabela 4. 10 expressões aspecto mais frequentes em cada categoria.

podemos concluir que os usuários costumam fazer menções diretas e indiretas aos atributos canônicos do produto usando várias expressões distintas. Este estudo contribui para a compreensão do impacto dos atributos canônicos do produto nas opiniões dos usuários e nossos resultados indicam que as opiniões dos usuários são realmente guiadas pelos atributos das especificações do produto e destacam a influência dos atributos canônicos nas avaliações dos usuários. Algumas limitações estão associadas a este estudo, as quais, entretanto, podem fornecer direcionamentos para pesquisas futuras. Primeiro, consideramos apenas expressões de aspecto para representar as opiniões do usuário. Pesquisas futuras podem estender o estudo atual para examinar outros componentes de opiniões, como avaliações numéricas (*ratings*), polaridade de opinião e tempo de postagem de opinião. Segundo, nossa análise é restrita a produtos que, por sua vez, possuem um conjunto bem estabelecido de atributos canônicos fornecidos pelos fabricantes. No entanto, domínios como restaurantes ou hotéis não têm atributos canônicos claros. Portanto, pesquisas futuras podem estender o estudo atual a esses domínios.

5. Trabalhos Relacionados

A discussão apresentada neste artigo está em linha com esforços recentes que visam estruturar conhecimento subjetivo sobre entidades [Halevy 2019]. O objetivo final é permitir que usuários e aplicações possam com lidar informações subjetivas de forma mais eficaz. Um desses esforços recentes é o *OpineDB* [Li et al. 2019], um sistema de banco de dados subjetivo que armazena opiniões extraídas de resenhas de usuários. Essas opiniões

são estruturadas de acordo com um esquema de banco de dados subjetivo e permite que consultas subjetivas sejam processadas sobre essas opiniões.

O *ModSpot* [Vieira et al. 2015] é um método semi-supervisionado projetado para aprender um modelo *Conditional Random Fields (CRF)* que identifica números de modelo de produto que ocorrem em sentenças extraídas de postagens de fóruns dadas como entrada. Para o processo de aprendizado, o método requer apenas um conjunto de exemplos de números de modelo de semente na mesma categoria, o que significa que ele não exige que sentenças de treinamento anotadas sejam fornecidas. O *ProdSpot* [Vieira 2018] expande o *ModSpot*,ao criar um sistema de reconhecimento de menções a produtos de alto desempenho que rotula menções de produto a partir de conteúdo gerado pelo usuário, tomando como entrada apenas descrições não-estruturadas de produtos e o texto gerado pelos usuários. Isto é conseguido sem quaisquer dados manualmente rotulados, através da inicialização de um classificador supervisionado usando um conjunto de exemplos de menções aos produtos extraídas automaticamente das das descrições dos produtos.

Outro método chamado *ProdLink* [Vieira et al. 2016] foi desenvolvido para vincular as menções reconhecidas à sua contraparte do mundo real. Os autores argumentam que esse problema pode ser efetivamente resolvido usando um conjunto de evidências que podem ser extraídas do conteúdo de mídia social e das descrições de produtos. Especificamente, mostramos quais características devem ser usadas, como elas podem ser extraídos e como combiná-las através de técnicas de aprendizado de máquina. O ProdLink é uma solução supervisionada para vinculação de produtos, capaz de reconhecer as menções de produtos em texto em linguagem natural a partir de publicações em fóruns públicos e de vincular as menções às entradas em um catálogo.

Mais relacionado com esse artigo, é uma nova abordagem proposta para enriquecer catálogos de produtos com opiniões de usuários no nível de granularidade de atributo. Nossa abordagem realiza duas tarefas distintas, mas relacionadas: identificar sentenças com opiniões direta nas comentários feitos por usuários e mapear estas opiniões para atributos de produtos em um catálogo. Para a primeira tarefa foram propostos dois métodos alternativos. O primeiro método, chamado AspectLink [de Melo et al. 2018], adota uma estratégia não-supervisionada, enquanto o segundo, chamado OpinionLink [de Melo et al. 2019a], adota uma estratégia supervisionada. O AspectLink [de Melo et al. 2018] usa uma abordagem linguística não supervisionada onde funções de similaridade comparam as características lexicais de atributos dos produto com o texto das expressões de aspecto. No OpinionLink [de Melo et al. 2019a], foi desenvolvido um método que usa classificadores binários e um conjunto de características estatísticos extraídas das opiniões de usuários. Além disso, uma estratégia de bootstrapping foi proposta para treinar os classificadores a fim de reduzir a dependência dos dados de treinamento.

Ainda neste tópico, foi realizado um estudo empírico sobre o uso de menções diretas e indiretas nas avaliações de usuários sobre os atributos de produtos. Os resultados deste estudo indicam que as opiniões dos usuários são guiadas pelos atributos de produtos e destacam a influência dos atributos nas opiniões dos usuários [de Melo et al. 2019b]. Os dados utilizados neste artigo são derivados deste estudo.

Referências

de Melo, T., da Silva, A. S., and de Moura, E. S. (2018). An aspect-driven method for enriching product catalogs with user opinions. *J. Braz. Comp. Soc.*, 24(1):15:1–15:19.

de Melo, T., da Silva, A. S., de Moura, E. S., and Calado, P. (2019a). *OpinionLink*: Leveraging user opinions for product catalog enrichment. *Inf. Process. Manage.*, 56(3):823–843.

de Melo, T., da Silva, A. S., de Moura, E. S., and Calado, P. (2019b). The importance of canonical product attributes on user opinions: an empirical investigation. In *10th Latin American Web Congress, LA WEB 2019*, pages 772–778.

Doan, A., Halevy, A., and Ives, Z. (2012). *Principles of data integration*. Elsevier.

Halevy, A. Y. (2019). The ubiquity of subjectivity. *IEEE Data Eng. Bull.*, 42(1):6–9.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Kostyra, D. S., Reiner, J., Natter, M., and Klapper, D. (2016). Decomposing the effects of online customer reviews on brand, price, and product attributes. *International Journal of Research in Marketing*, 33(1):11–26.

Li, Y., Feng, A., Li, J., Mumick, S., Halevy, A., Li, V., and Tan, W.-C. (2019). Subjective databases. *Proc. VLDB Endow.*, 12(11):1330–1343.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2:627–666.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*.

Pang, B. and Lee, L. (2007). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Park, E. J., Kim, E. Y., Funches, V. M., and Foxx, W. (2012). Apparel product attributes, web browsing, and e-impulse buying on shopping websites. *Journal of Business Research*, 65(11):1583–1589.

Poria, S., Cambria, E., Ku, L.-W., Gui, C., and Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, page 28.

Qadir, A. (2009). Detecting opinion sentences specific to product features in customer reviews using typed dependency relations. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 38–43. Association for Computational Linguistics.

Vieira, H. S. (2018). Recognition and linking of product mentions in user-generated contents. Programa de Pós-Graduação em Informática, Universidade Federal do Amazonas. Tese de Doutorado.

Vieira, H. S., da Silva, A. S., Calado, P., Cristo, M., and de Moura, E. S. (2016). Towards the effective linking of social media contents to products in e-commerce catalogs. In *CIKM*, pages 1049–1058. ACM.

Vieira, H. S., da Silva, A. S., Cristo, M., and de Moura, E. S. (2015). A self-training CRF method for recognizing product model mentions in web forums. In *37th European Conference on IR Research, ECIR 2015*, pages 257–264.

Altigran Soares da Silva é professor titular do Instituto de Computação da UFAM (IComp/UFAM). Concluiu seu doutorado pela UFMG em 2002. Tem coordenado e participado de dezenas de projetos que resultaram em mais de 130 publicações científicas. Foi coordenador de comitês de conferências no Brasil e no exterior e membro de comitês de programa em cerca de 50 conferências e workshops internacionais. Exerceu a Pró-reitoria de Pesquisa e Pós-Graduação da UFAM (2007/2009), foi Coordenador Adjunto da área de Computação na CAPES (2011/2013), membro do CA-CC do CNPq (2016/2019), membro da diretoria da SBC (2005/2015), e membro do conselho da Sociedade entre (2016/2019). É co-fundador de empreendimentos de tecnologia, como a Akwan (adquirida pela Google Inc.,2005), Neemu (adquirida pela Linx Sistemaa, 2015) e Teewa (adquirida pela JusBrasil, 2019). Recebeu o 1o. Lugar no Concurso de Teses e Dissertação da SBC (2013), Menção Honrosa no Prêmio CAPES de Teses (2013), prêmio de Sócio Destaque da SBC (2013) e Google Research Awards in Latin America (2015).