Uma Classificação Facetada baseada em Análise de Sentimento para Sistemas de Recuperação de Informação

***Resumo*—Presenciamos, no decorrer dos anos, um crescimento bastante significativo da quantidade de informações disponíveis às pessoas, principalmente no que diz respeito à *Web*. Desta forma, se por um lado este crescimento representa algo bom, por disponibilizar aos usuários mais meios de obter um determinado conhecimento, por outro, ele acarreta em um grande problema: determinar qual desses meios são os melhores, isto é, qual deles está mais próximo de solucionar o problema do usuário. Assim sendo, propomos neste trabalho utilizar a análise de sentimento como uma maneira diferente de apresentar, às pessoas, os resultados retornados por um Sistema de Recuperação de Informação (SRI) qualquer. A aplicação da análise de sentimento tem o objetivo de agrupar os documentos de acordo com suas polaridades. Então, os usuários teriam a opção de usar a ferramenta para filtrar a lista de resultados, analisando apenas aqueles documentos que realmente fossem interessantes. Ou seja, visamos realizar uma classificação facetada com a análise de sentimento. Vale ressaltar, ainda, que os sentimentos são determinados pelas sentenças opinativas contidas nos documentos. Com o intuito de exemplicar toda essa teoria fizemos uma demonstração de como ela seria na prática.**

Palavras-chave - Análise de sentimento; mineração de opinião; extração de opinião; recuperação de informação; classificação facetada; sentiment analysis; opinion mining; opinion extraction; information retrival; faceted classification

I. Introdução

A busca por respostas sempre esteve atrelado à existência humana. É da natureza do homem esta "sede" por informações que supram suas dúvidas ou ajudem a solucionar seus problemas. Logo, quando uma pessoa se depara com situações que exigem um determinado conhecimento, o qual ela ainda não adquiriu ou apenas não aperfeiçoou, a primeira atitude tomada é procurar por fontes que auxiliem na obtenção deste. A área de Recuperação de Informação (RI) envolve estudos que visam melhorar os meios, e os resultados, nesse esforço de alcançar novos conhecimentos [1].

No entanto, atualmente, é possível encontrar milhares de fontes de informações para uma única dúvida. Um bom exemplo da vasta quantidade de materiais, contendo dados com respostas a diversos obstáculos, é a *Web*. Nela, realmente, ao procurarmos por uma determinada informação recebemos como resultado milhares de locais nos quais, supostamente, possuem o que de fato desejamos. Porém, tais resultados podem conter informações próximas, ou até mesmo, informações totalmente sem nexo, daquilo que se almejava encontrar. Essa situação acarreta no seguinte problema: existe um número muito grande de materiais, tornando inviável a atividade do usuário verificar um por um. Resolver este problema é o principal objetivo da RI. Isto é, o objetivo fim da recuperação de informação é selecionar os materiais que realmente contenham as informações relevantes, das quais o usuário necessita, esforçando-se para classificá-los de acordo com a "importância" de cada um individualmente [2].

Existem muitas discussões em torno de técnicas e abordagens que buscam melhorar os resultados obtidos por Sistemas de Recuperação de Informação (SRI). Neste trabalho, não estamos focados na extração da informação ou na relevância dos materiais retornados, mas sim numa maneira de auxiliar o usuário na filtragem desses materiais. Para isso usamos um método chamado Análise de Sentimento (AS), o qual procura extrair do texto a atitude transpassada pelo autor. Ou seja, a AS tem como propósito determinar se um dado texto exprime um "sentimento" positivo, negativo ou, simplesmente, neutro. Consequentemente, nós propomos utilizar a Análise de Sentimento com a finalidade de elaborar uma ideia sobre uma ferramenta de apoio à classificação das fontes de informações. Isto é, o resultado de uma determinada busca por respostas, na maior parte das vezes, retorna muitos materiais - como dito no segundo parágrafo, isto posto, utiliza-se a AS para apresentar ao usuário uma nova forma de reordenação, usando a classificação referente a polaridade - positivo ou negativo - dos documentos.

Em suma, estamos trantando de uma classificação conhecida como classificação facetada, uma técnica cujo intuito é colocar os materiais que tratam de um mesmo assunto em conjuntos [3] [4]. Visto que, nossos estudos estão focados em classificar uma lista de documentos, em ordem do mais positivo para o mais negativo ou o contrário. Porém, vale ressaltar que essa ordem é determinada pela quantidade de sentenças positivas e/ou negativas, presentes em cada documento analisado. Ou seja, caso deseja-se classificar do mais positivo para o mais negativo, então o documento presente na primeira posição será aquele cujo número de sentenças positivas seja o maior entre todos os avaliados. Ainda, dentro dos documentos é feito uma marcação das sentenças encontradas, distinguindo as positivas das negativas.

Nosso artigo está estruturado de uma maneira que auxilie a compreensão individual de cada tema envolvido, evoluindo para o entendimento de o porquê de combiná-los. Ademais, apresentamos alguns trabalhos que possuem características semelhantes ao nosso e ideias que podem vir a ser implementadas. Sendo

assim, tem-se na Seção Trabalhos Relacionados estudos cujos perfis são similares aos nossos e que formaram o alicerce desses, na Seção de Recuperação de Informação definições e exemplos referentes a RI, bem como alguns conceitos sobre classificação facetada, na Seção Análise de Sentimentos descrevemos a teoria referente à mineração de opinião e quais os motivos de utilizá-la em nosso trabalho, na Seção da Metodologia apresentamos a proposta deste trabalho e os detalhes que nos levaram a estudá-la, além, é claro, da demonstração, na Seção de Conclusão o fechamento da proposta do artigo em questão, e, por fim, na Seção de Trabalhos Futuros possíveis implementações que poderiam ocorrer.

II. Trabalhos Relacionados

O foco do nosso trabalho está, principalmente, na tentativa de combinar os recursos da análise de sentimento e de um sistema de recuperação de informação, com a finalidade de melhorar a maneira como é apresentada a lista de resultados ao usuário. E, foi a partir desta forma de pensar que encontramos a classificação facetada. Visto que, seu propósito é agrupar materiais cujo conteúdo seja o mesmo - no nosso caso a polaridade. Para seguir com este raciocínio identificamos alguns trabalhos com ideias similares às nossas.

Por exemplo, Zheng e Fang [5] propuseram um sistema que "retorna" documentos com base em seus sentimentos. Tal sistema faz a tarefa de recuperação de informação, mas, além disso, verifica o sentimento expresso em cada documento individualmente. Assim, os resultados obtidos são classificados entre documentos "contra" (negativos) e documentos "a favor" (positivos). Além disso, o sistema associa documentos que tratam do mesmo tema e pertencem a mesma categoria - positivo ou negativo. Zheng e Fang [5] citam que através desse sistema o usuário pode, facilmente, identificar e comparar argumentos positivos e argumentos negativos, sem a necessidade de ler todos documentos retornados. Esta ideia é bem próxima da nossa proposta. Contudo, há uma diferença marcante entre as duas. Enquanto Zheng e Fang [5] fazem uma análise de sentimento em nível de documento, nós nos aprofundamos numa análise em nível de sentença. Isto é, no primeiro caso, o sentimento retornado é obtido a partir de uma análise mais geral do documento em questão, já no segundo, tal sentimento refere-se a um conjunto de análises de cada sentença presente nele. Desta forma, podemos também realizar diversos tratamentos para identificar as opiniões que, de fato, referem-se a entidade avaliada. Tais tratamentos são descritos na Seção "Metodologia".

Ademais, alguns trabalhos demonstram como a classificação facetada pode facilitar essa ligação, entre a análise de sentimento e a recuperação de informação. Tal como, Broughton [6], o qual propõe analisar o impacto da classificação facetada sobre ferramentas de recuperação de informação. O trabalho de Broughton [6] foi verificar o desempenho de vários sistemas de recuperação de informação utilizando uma classificação multi-facetada. Mills [7] também visa, em seus estudos, combinar a classificação facetada com a RI. Porém, Mills [7], ao invés de analisar o desempenho desta combinação, apresenta os detalhes do papel que essa classificação possui na recuperação de informação. Por fim, existem alguns trabalhos, como o demonstrado por Prieto [8], que aplicam essa combinação, de classificação facetada com recuperação de informação, na prática de fato, exibindo sua performance.

Todos esses trabalhos, citados nos parágrafos anteriores, formaram a base dos nossos estudos. Ou seja, os problemas apresentados pelos autores nos impulsionaram até o foco deste trabalho: elaborar uma ideia que auxilie na identificação das respostas de um determinado problema, por meio de uma classificação facetada. Desta maneira, conseguimos elaborar um alicerce muito bem estruturado por meio de trabalhos já comprovados anteriormente.

III. Recuperação de Informação

Quando uma pessoa qualquer encontra-se numa determinada situação que exige um novo conhecimento, comumente, ela utilizará um meio de recuperação de informação para obtê-lo. Visto que, podemos dizer que Recuperação de Informação (RI) é a atividade de buscar por um novo conhecimento no momento que surge uma dada necessidade de informação. Apesar da RI ser uma tarefa multidisciplinar, ela é considerada um campo de estudo da área de Informática, pois a maioria das fontes de informações, atualmente, estão presentes em sistemas computacionais. Cardoso [9] trata a Recuperação de Informação como uma técnica de armazenamento automático e recuperação de documentos.

Segundo Manning et al [1], Recuperação de Informação é a ação de encontrar um material, geralmente representado por documentos, cuja natureza seja não-estrutural - uma forma que não seja reconhecida naturalmente pelo computador, mas que satisfaça uma necessidade de informação. Na maior parte das vezes, este material pertence a uma grande coleção de materiais, o que torna a tarefa de encontrá-lo um procedimento árduo.

O objetivo dos estudos atrelados a Recuperação de Informação é fazer com que os resultados - documentos recuperados - se aproximem cada vez mais daquilo que o usuário almeja encontrar. Isto é, aperfeiçoar as técnicas de recuperação, de modo que estas consigam identificar quais os materiais que, realmente, são relevantes para a necessidade apresentada pelo usuário em questão. Ademais, o modo como tais fontes são apresentadas também possuem uma grande influência no pontecial de pesquisa do usuário [10], pois dependendo da maneira que se dá a apresentação, é possível elimininar materias que não interessem às pessoas.

Certas situações, como a apresentada no primeiro parágrafo, nos levam a recorrer ao uso de atividades relacionadas a RI. Tais atividades, normalmente, são agrupadas e automatizadas, recebendo o nome de Sistema de Recuperação de Informação (SRI). Diversas vezes nos apropriamos desses sistemas sem ter consciência de que estamos fazendo. Como exemplos de SRIs tem-se os motores de busca da *Web*, consultas eletrônicas em bibliotecas, ou até mesmo pesquisas realizadas nos sistemas operacionais, nos quais trabalhamos. A Figura 1 [9] ilustra as estapas básicas que compõem um SRI.

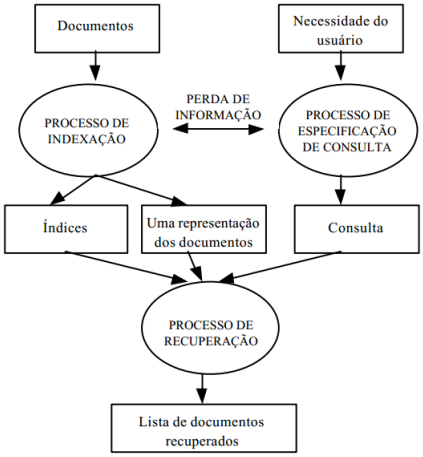


Figura 1. Componentes básicos de um Sistema de Recuperação de Informação

Note que o padrão dos sistemas de recuperação de informação possui o mesmo objetivo que a base da RI. Primeiramente, temos a necessidade do usuário e os documentos, já existentes no domínio do SRI. Em seguida, ocorrem diversos processos com o intuito de selecionar os documentos mais próximos da solução, para tal necessidade. Por fim, são listados, ou melhor, apresentados os documentos resultantes do processo citado anteriormente.

Nos dias atuais, com o crescimento demasiado da quantidade de novas informações eletrônicas, cresce também a necessidade de técnicas mais apuradas para sistemas de recuperação de informação [2]. Visando proporcionar uma situação melhor para tal problema, nosso trabalho propõe utilizar a Análise de Sentimentos como uma ferramenta de auxílio para a apresentação dos documentos recuperados, separando para o usuário os documentos positivos e os negativos referentes a sua necessidade de informação.

Logo, podemos dizer que nossos estudos estão focados sobre a forma como os documentos são apresentados para o usuário. Isto é, o propósito do nosso trabalho refere-se a uma nova maneira de organizar a lista de documentos recuperados - a última componente do processo de SRI. Segundo Hearst [10], quando a visualização de informação se baseia apenas em textos torna-se mais dificultoso a tarefa de identificar, nos documentos recuperados, quais aqueles que contém a melhor resposta para a dúvida em questão. Pensando nisso, propusemos uma classificação baseada no sentimento de cada documento, pois dessa forma, o usuário pode escolher entre aqueles que trazem uma informação positiva e as que contém uma informação negativa, sobre o assunto.

A. *Classificação facetada*

A classificação facetada é uma forma de agrupar materiais que tenham uma determinada relação, por meio de suas características ou ações. Isto é, identificar elementos que pertencem a um mesmo conjunto, quando trata-se de um dado aspecto ou um conceito qualquer [3]. Segundo Barbosa [11], a análise facetada coordena conceitos. Por conta disso, ele aponta que até mesmo assuntos muito complexos podem ser representados por uma classificação facetada.

Neste trabalho, pretendemos elaborar uma ideia que utilize a classificação facetada, juntamente com a análise de sentimento, para reorganizar a lista de resultados, de um determinado SRI, de acordo com a polaridade das opiniões encontradas. Ou seja, nossa proposta é de uma ferramenta de apoio à pesquisa dos usuários. Sendo assim, um usuário qualquer pode, no momento de sua pesquisa, classificar os documentos retornados de acordo com o sentimento presente em cada um deles. Essa classificação pode ser feita do documento mais positivo para mais negativo, ou do mais negativo para o mais positivo - através da análise das sentenças opinativas presentes neles. Pretendemos, desta forma, diminuir a quantidade de materiais que as pessoas precisam ler para encontrar as respostas, aos seus problemas.

IV. Análise de Sentimento

O foco principal da análise de sentimento são as opiniões. É a partir delas que as pessoas expressam seus sentimentos referentes a uma determinada entidade. Tais sentimentos podem ser classificados como positivos, as pessoas estão, pelo menos, satisfeitas com a entidade avaliada, ou negativos, quando encontra-se opiniões as quais manifestam desinteresse ou desapresso para com a entidade em questão [12]. Além disso, podemos encontrar um sentimento neutro. Para um sentimento ser categorizado como neutro significa que não existe opinião no texto avaliado, mas sim um fato. De um modo geral, a ação de realizar uma análise de sentimento pode ser descrita como classificar textos de acordo com a opinião contida nele.

Segundo Bing Liu [13], as opiniões estão no centro das atividades humanas, por que elas representam a chave que influencia os nossos comportamentos. Por exemplo, enquanto de um lado tem-se as empresas que precisam descobrir quais as opiniões dos consumidores, em relação a seus produtos, permitindo assim que elas planejem sua estratégia de vendas. Do outro estão os consumidores, os quais, constantemente, procuram por informações sobre determinados produtos que lhes interessam. Dessa forma, podem reunir dados o suficiente para sanar suas dúvidas, e ajudar à tomar a decisão de comprar ou não aquele produto.

Embora seja uma área de pesquisa recente, tendo início por volta dos anos 2000, a análise de sentimentos está crescendo constantemente, e é um campo de estudos muito ativo. Pois, para qualquer domínio dependente da opinião das pessoas, ela se torna extremamente útil. A importância dessa análise tomou proporções tão extensas que despertou o interesse de grandes organizações, tais como Microsoft, Google, Hewlett-Packard, entre outras [13].

Não é uma tarefa trivial realizar a leitura e o entendimento do grande volume de opiniões - atreladas a uma determinada entidade - que encontramos hoje em dia. Entretanto, Bing Liu [13] cita três motivos que impulsionam os estudos na área de análise de sentimento. Primeiro, o grande interesse das organizações, em utilizar tal análise em aplicações comerciais, geram uma forte motivação para pesquisa. Segundo, ainda existem muitos problemas sem solução, ou seja, problemas

que nunca foram estudados antes. Terceiro, o vasto volume de opiniões nas mídias sociais, principalmente na *Web*, é algo que não existia no passado, e que pode possibilitar outros tipos de pesquisas. Ademais, apesar da análise de sentimento ser um campo de estudos do Processamento de Linguagem Natural (PLN), ela pode ter um impacto intenso em outras áreas, como por exemplo, na política, na economia e na administração.

A classificação de documentos quanto o sentimento encontrado pode se tornar uma ferramenta de apoio à apresentação de informações de grande potencial, no que diz respeito a recuperação de informação. Visto que, os usuários de RI teriam a possibilidade de classificar os resultados de uma determinada pesquisa mediante as opiniões extraídas desses. Isto posto, essa situação poderia reduzir os "alvos" a serem analisados pelos usuários.

A. Processo da Análise de Sentimento

Atualmente, existem inúmeros diferentes métodos que realizam análise de sentimento. No entanto, apesar, de cada método possuir suas particularidades, há um fator em comum em todos: a base da análise de sentimento. Isto é, mesmo podendo ter diferenças gritantes, todos os métodos envolvidos nesta área estão alicerçados pelo mesmo processo básico de análise de sentimento. Tal processo é composto pelas seguintes etapas: extração de opinião, identificação dos sentimentos e sumarização dos resultados. A Figura 2 ilustra um fluxograma da ocorrência desse processo [13] [14].

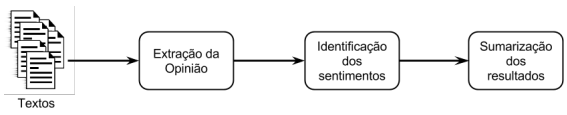


Figura 2. Etapas básicas para a análise de sentimento

A etapa de extração de opinião visa procurar e retirar os comentários contidos no texto em análise. Enquanto que, na etapa seguinte, o foco já está sobre cada palavra presente no conteúdo extraído. A fase de identificação dos sentimentos verifica qual o sentimento que os termos expressam individualmente. Podemos dizer que, dentre as três etapas básicas, esta é a mais importante no processo de análise de sentimento. Por último, temos a sumarização dos resultados, a qual representa a etapa de exposição dos valores obtidos.

Nosso trabalho, como todos os outros, utilizou este processo, apresentado pela Figura 2, como base. Nosso foco está na extração dos comentários, pois buscamos encontrar aqueles que realmente tratam da entidade analisada, e na última etapa, a qual procuramos combinar com os resultados obtidos pela RI.

B. Níveis de Análise de Sentimento

O termo granularidade, empregado na análise de sentimento, é utilizado para descrever o grau de detalhes presente num determinado método. Em geral, existem três níveis nos quais um ou mais métodos podem se aprofundar: nível de documento, nível de sentença e nível de entidade e aspecto [15]. Na Figura 3 demonstramos como esses níveis estão dispostos no texto em questão. Em cada um deles há um foco diferente. Isto é, cada nível possui suas próprias características e particularidades. Logo, quando deseja-se aplicar um método de análise de sentimento é necessário ter em mente a direção correta que pretende-se seguir. Ou seja, é preciso, em primeiro lugar, formular a abordagem que espera-se atingir.

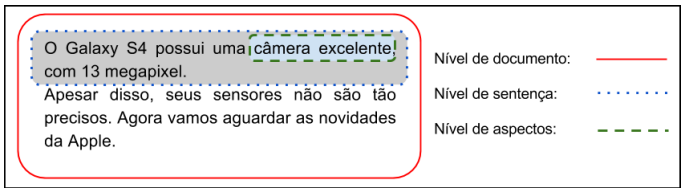


Figura 3. Detalhamento dos níveis de análise de sentimento

1) Nível de Documento: A análise de sentimento em nível de documento, basicamente, tem o propósito de verificar se o sentimento, expresso pela opinião, de um determinado documento é positivo ou negativo, em relação a uma dada entidade [16], [17]. Por exemplo, supomos que almeja-se encontrar o tipo de opinião para o produto X, e tem-se apenas um documento, contendo a revisão de tal produto. Então, a análise de sentimento irá averiguar qual a opinião expressa no documento em geral. Desta forma, o resultado obtido pela análise será a opinião sobre o produto X. Entretanto, neste nível é necessário assumir que cada documento investigado apresenta um sentimento referente a uma única entidade, tornando-o um tanto limitado.

2) Nível de Sentença: A análise de sentimento em nível de sentença procura identificar se uma determinada sentença possui sentimento positivo, negativo ou neutro, referente a entidade avaliada [13]. Como exemplo, tomemos a seguinte frase: *"Eu gostei bastante do carro"*. Caso fosse feito uma análise, o resultado seria um sentimento positivo. No entanto, podem haver sentenças que "confundam" a análise de sentimento neste nível, tais como, *"Os pneus do carro estão mais que perfeitos, mas o freio está falhando"*. Nesta frase, existe o "lado bom e o ruim". Porém, a análise de sentimento em nível de sentença não tem "poder" para discernir este problema.

Embora haja essa limitação, este nível de análise é mais do que necessário para o nosso trabalho. Visto que, nosso propósito é verificar as opiniões, contida em cada documento, presente na lista de resultados da recuperação de informação, sobre a pesquisa realizada pelo usuário. Sendo que, tal verificação se dá pela quantidade de sentenças opinativas inclusas nos documentos analisados.

C. Tipos de Opiniões

Segundo Bing Liu [13], [18], [19] podemos classificar as opiniões em dois grandes tipos: opiniões regulares e opiniões comparativas. No primeiro, temos aquelas opiniões mais simples e que tratam apenas de uma única entidade. Ou seja, quer a análise de sentimento aconteça em nível de documento quer em nível de sentença, ou ainda, em nível de entidade e aspecto, ela sempre irá tratar somente de uma entidade,

ignorando a presença de outras, quando encontradas nos textos analisados. Já no segundo, encontram-se as opiniões um tanto mais complexas. Visto que, nesse caso considera-se a presença de mais do que uma entidade. Isto é, o texto em questão pode conter múltiplas entidades, e a análise de sentimento precisa distinguir as opinião relacionadas a cada uma para poder compará-las.

Neste trabalho optamos por tratar, exclusivamente, das opiniões regulares, pois o foco inicial é distinguir documentos negativos de documentos positivos referentes a uma dada entidade, presente na pesquisa de um determinado usuário. Sendo assim, as opiniões sobre outras entidades não serão consideradas.

Ademais, utilizamos algumas técnicas de extração de opinião para diferenciar os comentários alusivos a entidade em análise, de outras entidades presentes no texto. Desta forma, tomamos para a análise de sentimento somente aquelas sentenças que realmente fazem alusão a entidade em questão. Logo, percebe-se que não existe comparação entre as entidades, mas sim uma distinção de qual é a entidade principal e quais são as entidades secundárias.

D. RI na Análise de Sentimento

Existem diversos ramos na recuperação de informação e na análise de sentimento que acabam concorrendo a um mesmo fim. Isto é, há campos, nessas duas áreas da informática, que visam um bem em comum. Como exemplo disso temos os estudos referentes à *Opinion search*, os quais, de acordo com Bing Liu [12], nada mais são do que uma combinação da recuperação de informação e da análise de sentimento.

O processo da *Opinion search* segue uma sequência similar à que ocorre na recuperação de informação. Primeiro recupera-se os documentos, relevantes à necessidade de informação do usuário, e, em seguida, faz-se o ranqueamento desses documentos. No entanto, após identificar os documentos relevantes à pesquisa do usuário, determina-se o sentimento, sobre a entidade ou o aspecto de uma entidade, presentes na *query*, dos resultados recuperados. Isto é, cada material presente na lista de resultado recebe um sentimento positivo ou um sentimento negativo [13] [20].

Vale ressaltar que os materiais recuperados são classificados em: documentos contendo opiniões e documentos que não as contém. Os documentos com opinião ainda sofrem uma segunda classificação: positivos, negativos e mistos. Ou seja, identifica-se os documentos cujos sentimentos são positivos em um grupo, os com sentimentos negativos num segundo grupo, e, por fim, os documentos com opiniões mistas, nos quais encontra-se tanto sentimentos negativos quanto positivos, em um terceiro grupo [21].

Portanto, podemos observar que a ideia contida neste trabalho possui algumas semelhanças com a chamada *Opinion search*. Visto que, nosso intuito é, posteriormente à etapa de recuperação dos documentos, aplicar a análise de sentimento, de forma que o usuário possa utilizar tal identificação como uma ferramenta de apoio a sua pesquisa. Porém, é importante destacar que nossos estudos não visam documentos com sentimento misto. Ou seja, trabalhamos apenas com sentimento positivo ou negativo.

V. METODOLOGIA

Segundo Hearst [10], as pessoas têm uma maior facilidade de percepção por meio de imagens e representações visuais - quando comparados a formas, simplesmente, escritas. Por conta disso, nos focamos em encontrar uma maneira de tornar as informações mais fáceis de serem assimiladas pelos usuários, mesmo quando elas estão dispostas em textos. Hoje em dia, existe uma quantidade de fontes de informações muito grande, tal como Kobayashi e Takeda [22] apresentam em suas pesquisas. Seguindo este raciocínio, ferramentas de apoio à visualização de resultados seriam algo bastante importantes. No entanto, de acordo com o trabalho realizado por Chen e Yu [23], a visualização não influencia no desempenho de uma determinada pesquisa. Eles conduziram uma meta-análise em estudos de usabilidade de visualização de informação, e descobriram que o efeito gerado pela visualização não foi estatisticamente significativo.

Embora haja esta discussão, se existe, ou não, influência da visualização de informação na atividade de pesquisa das pessoas, só temos a certeza de uma única situação: os usuários estão acostumados às listas de materias, retornadas após uma determinada pesquisa. Ou seja, já existe um paradigma implantado na sociedade: o uso de lista de resultados para buscas. Sendo assim, propomos a ideia de uma ferramenta de apoio à apresentação da informação, continuando o aproveitamento das lista de resultados.

Nesse sentido, chegamos a conclusão que poderíamos aplicar a análise de sentimento sobre cada documento, retornado por um sistema de recuperação de informação qualquer, com a finalidade de agrupá-los em conjuntos cuja polaridade fosse a mesma. Por exemplo, teríamos um conjunto contendo documentos apenas positivos, outro somente negativos, e, por fim, um conjunto com os documentos neutros. Ainda, gostaríamos que nossa proposta fosse, realmente, apenas uma ferramenta de apoio. Ou seja, o usuário que almejasse utilizá-la, no intuito de encontrar suas respostas mais rapidamente, assim o faria. Caso contrário basta não usá-la. Visto que, em muitos casos, dependendo do tipo de busca, esta ideia poderia ser algo inútil, ou sem muito "valor", então o usuário poderia optar por não aproveitá-la.

Logo, ao reunirmos tais ideias, podemos perceber que esse trabalho visa realizar uma classificação facetada, demonstrando quais os documentos positivos, quais os documentos negativos e quais o documentos sem opiniões - neutros. Sendo assim, o usuário pode classificar a lista de resultados a partir do material mais positivo ou o mais negativo, entre os materiais retornados. Em suma, estamos combinando três campos diferentes: a recuperação de informação, com a finalidade de buscar por uma determinada necessidade de informação e retornar uma lista de resultados, a análise de sentimento, a qual verificará as opiniões contidas em cada documento da lista de resultados, por meio da identificação das sentenças opinativas, e a classificação

facetada, que permitirá ao usuário reordenar a lista de resultados de acordo com o sentimento que deseja, positivo ou negativo.

Para demonstrarmos essa nova técnica aplicamos alguns testes sobre uma amostra de documentos. Tal amostra foi obtida a partir de um SRI, e retornada em forma de lista de resultados. Em seguida, submetemos cada documento à análise de sentimento, conseguindo assim o sentimento expresso, individualmente, neles, por meio das sentenças opinativas identificadas. Daí, se fez necessário apenas agrupá-los entre os similares, quanto à polaridade, tornando possível a classificação facetada. A demonstração é apresentada e conceitualizada, de forma detalhada, nas seções seguintes.

A. Extração das Opiniões

A maioria das pesquisas em análise de sentimentos são aplicadas sobre comentários de produtos e, geralmente, os textos desta natureza fazem referência a somente um produto, expressam apenas uma única opinião e possuem pouco conteúdo irrelevante - que podem prejudicar a análise [24]–[26]. Contudo, neste trabalho pretendemos analisar notícias - e não comentários, extraídas de um determinado sistema de recuperação de informação. Nas notícias encontramos uma grande quantidade de opiniões, positivas e/ou negativas. Porém, diferente dos comentários encontrados nas redes sociais, esses conteúdos, comumente, estão escritos em uma linguagem mais formal e se encontram em fontes de informações sérias, possuindo assim um maior grau de confiabilidade.

Entretanto, para aplicações deste tipo, não é interessante executar uma análise de sentimento em nível de documento, pois um único texto pode conter vários tipos de opiniões, sejam elas positivas ou negativas. Também, o autor do texto pode introduzir comentários e opiniões referentes a várias entidades, os quais podem ou não conter palavras consideradas opinativas, e que não expressam nada a respeito da entidade analisada. Neste caso, se torna importante o uso de uma granularidade mais fina. Ou seja, analisar cada sentença que compõe o texto, e que estão relacionadas à entidade de interesse. Com o intuito de ilustrar melhor o problema, vamos considerar o seguinte texto:

"Na semana passada comprei a câmera X, e gostei muito (1). A qualidade das fotos são boas (2). Achei ela bem leve, algo que me agrada bastante (3). Eu tinha uma câmera Y, porém ela é muito ruim, a bateria descarrega rápido demais, e a imagem é péssima(4)."

Figura 4. Um dos materias retornados na lista de resultados do Google, acrescido dos detalhes da análise de sentimento

O texto apresentado carateriza-se como um comentário, tal como reviews de produtos. Desta forma, se o interesse é pela opinião sobre a câmera X, então tem-se somente a primeira frase se referindo explicitamente a entidade em questão. Contudo após verificar com mais detalhes, podemos observar que as frases 2 e 3, são opinativas, e também são alusivas a câmera X. Sendo assim, existe um problema de atribuição de entidade, no qual é necessário verificar, em todo o documento, a ocorrência de frases opinativas atreladas a entidade avaliada. A quarta frase possui características opinativas, mas é necessário descartá-la, pois ela não faz referência a câmera X. Esse é um dos pontos que esse trabalho irá atacar: extrair do texto somente as sentenças opinativas, e que fazem menção a entidade de interesse do usuário.

De acordo com o trabalho apresentado por Balaur [27], a correferência anafórica pode ser a solução para parte desse problema – apresentado no parágrafo anterior. Segundo ela, na teoria, a correferência anafórica pode melhorar o desempenho dos sistemas de mineração de opinião. Todavia, no experimento conduzido por Balaur ficou evidente que quando tal recurso foi utilizado, o desempenho do sistema diminuiu. Jagtap [28] relata em seus estudos outra solução que pode ser empregada na mineração de opinião: a correferência pronominal. Porém, ele propõe essa aplicação apenas como um trabalho futuro. E, em alguns casos, isso pode não ser o suficiente. Visto que, o autor do comentário pode optar pela não utilização de pronomes no momento de direcionar a fala para a entidade. Por exemplo, na segunda frase do trecho citado acima, o autor expõe sua opinião sobre uma característica da câmera X, mas não faz menção a ela explicitamente, nem utiliza recursos linguísticos para tal referência.

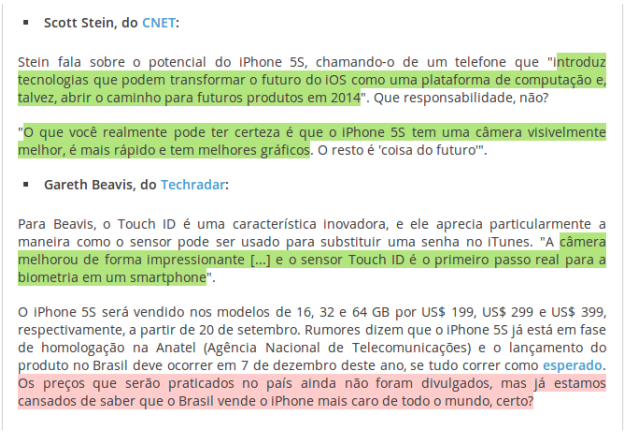
Neste trabalho, pretendemos testar algumas proposta de resolução de correferência, disponíveis na literatura, com a finalidade de buscar a melhor solução para a extração de opiniões. Ademais, de acordo com a definição de opinião apresentada por Bing Liu [13], além da entidade, alguns outros fatores podem estar envolvidos, como por exemplo, a data da publicação, nome do autor, o sentimento expresso no momento, entre outros.

Juntamente com a atribuição de entidade a frases, este trabalho irá abordar outros pontos, os quais segundo Bing Liu [13] são essenciais. Para tanto vamos considerar a notícia a seguir:

> 24/11/2013 09h55 - Atualizado em 24/11/2013 12h02 Após frase de Júlio Baptista, STJD analisará jogo entre Vasco e Cruzeiro O procurador-geral do STJD, Paulo Schmitt, não acredita que a afirmação de Júlio indique uma possível facilitação do Cruzeiro, porém prometeu investigar o caso (1).
> - Em tese, esse tipo de declaração/afirmação, de forma isolada e fora de contexto, poderia sugerir uma entrega ou manipulação de resultado de partida (2). Tal conduta, se confirmada, configuraria infração aos artigos 243 e 243-A do CBJD (3).
> É improvável que isso tenha ocorrido, pois ninguém revelaria de forma tão inocente um esquema que prejudicaria vários outros clubes em benefício de um deles (4). De todo modo, como pouca coisa me surpreende ultimamente, iremos avaliar imagens e lances do jogo, que é mais importante nesse

cenário do que qualquer diálogo ou provocação (5). E, caso haja facilitação, omissão em disputa, entre outros aspectos de impacto no resultado da partida, poderemos oferecer denúncia ou ao menos requerer uma investigação - afirmou Schmitt (6).
Perguntado sobre quem seria punido - Júlio Baptista isoladamente ou o Cruzeiro como um todo - se caso fosse comprovado qualquer tipo de manobra, Paulo reiterou não acreditar em má-fé por parte de nenhum dos envolvidos na vitória vascaína por 2 a 1 (7).

Um ponto importante que pretendemos trabalhar, e que vai de encontro com a definição de opinião apresentada por Bing Liu [13], é a data em que o comentário foi publicado. A data é interessante porque, na maioria dos casos, os jornais publicam algo que foi dito a muito tempo, e que pode não refletir mais uma situação atual. Ainda vale ressaltar que, muitas vezes, os comentários são escritos em datas diferentes daquela em que a notícia fora publicada. Até o momento não foi encontrado nenhum trabalho que abordasse esse ponto, por isso tentaremos propor algo que possa vir a ser uma solução.


- **Scott Stein, do CNET:**

Stein fala sobre o potencial do iPhone 5S, chamando-o de um telefone que "introduz tecnologias que podem transformar o futuro do iOS como uma plataforma de computação e, talvez, abrir o caminho para futuros produtos em 2014". Que responsabilidade, não?

"O que você realmente pode ter certeza é que o iPhone 5S tem uma câmera visivelmente melhor, é mais rápido e tem melhores gráficos. O resto é 'coisa do futuro'".

- **Gareth Beavis, do Techradar:**

Para Beavis, o Touch ID é uma característica inovadora, e ele aprecia particularmente a maneira como o sensor pode ser usado para substituir uma senha no iTunes. "A câmera melhorou de forma impressionante [...] e o sensor Touch ID é o primeiro passo real para a biometria em um smartphone".

O iPhone 5S será vendido nos modelos de 16, 32 e 64 GB por US$ 199, US$ 299 e US$ 399, respectivamente, a partir de 20 de setembro. Rumores dizem que o iPhone 5S já está em fase de homologação na Anatel (Agência Nacional de Telecomunicações) e o lançamento do produto no Brasil deve ocorrer em 7 de dezembro deste ano, se tudo correr como esperado. Os preços que serão praticados no país ainda não foram divulgados, mas já estamos cansados de saber que o Brasil vende o IPhone mais caro de todo o mundo, certo?

Figura 5. Documento escolhido para a demonstração das marcações feitas nas sentenças opinativas

Além de tentar identificar os comentários que pertencem a entidade de interesse e a data em que foi citado o comentário, também é importante verificar o autor de cada comentário. O autor do comentário nem sempre é o autor da notícia. Observando o trecho da notícia acima, podemos identificar duas pessoas, porém, é o procurador-geral do STJD que expõe o comentário, o qual exerce influência sobre as entidades Cruzeiro e Vasco. Encontramos alguns aplicativos que fazem a extração do autor, tal como o *AlchemyAPI*. Nossa proposta é buscar os comentários e, se possível, encontrar também o seu autor. Para ilustrar melhor, vamos considerar o trecho da notícia acima e o Cruzeiro como a entidade principal. Sendo assim, seriam recuperadas todas as sentenças que caraterizam comentário, a respeito do Cruzeiro, as frases de 2 a 6, e o autor das mesmas, Paulo Schmitt.

Ainda, é importante destacar que propomos marcar dentro do documento quais são os trechos que representam os sentimentos positivos ou negativos sobre a entidade de interesse. Outra proposta, também ligada a este trabalho, é acrescentar, por exemplo, ao *snippets* do Google, a quantidade de sentenças positivas e negativas presentes em cada documento, e se o documento se caracteriza como positivo ou negativo.

B. Demonstração

Com o intuito de apresentarmos nossa proposta na prática, fizemos uma demonstração de como ocorreria todo o processo da análise de sentimento sobre a recuperação de informação, por meio da classificação facetada. Como citado nas seções anteriores, primeiramente, um determinado sistema de recuperação de informação retorna uma lista de documentos, conforme a busca realizada pelo usuário. A busca que realizamos para a demonstração foi referente à "reviews Iphone 5S", ou seja, essa foi a *query* utilizada. Dentre todos os materiais retornados escolhemos apenas um para exibir as sentenças positivas e negativas presentes nele - mostrando a marcação dessas ao abri-lo - e como seria o *snippet* do Google. A Figura 4 ilustra os detalhes apresentados aos usuários na lista de resultados, e que auxiliariam na escolha de quais os documentos mais importantes para analisar de fato.

Já a Figura 5 apresenta parte do documento escolhido e algumas das sentenças opinativas marcadas, demonstrando a facilidade que seria para identificar aquela informação que o usuário deseja realmente.

Ainda, é importante ressaltar que, para esta demonstração, foi considerado que o usuário utilizou a classificação facetada com o intuito de expor a lista de resultados do documento mais positivo para o mais negativo.

VI. Conclusão

Em meio aos nossos estudos pudemos perceber que, de fato, existe uma relação entre as áreas de análise de sentimento e recuperação de informação. Diversos trabalhos e conceitos tratam dos detalhes deste elo, tais como Zheng e Fang [5], ao descreverem as particularidades de seu sistema, e Bing Liu [13], relatando acerca de *Opinion Search*. Ademais, descobrimos que essa relação pode ter duas ramificações distintas: uma focada na análise de sentimento, tendo apenas a RI como uma ferramenta de apoio, e outra com o foco sobre a recuperação de informação, sendo a análise de sentimento somente algo complementar. No primeiro caso, tem-se o próprio processo da mineração de opinião como exemplo. Visto que, a etapa de extração de opinião utiliza muitos conceitos da RI. Contudo, o ponto central deste trabalho "pairou" acima da segunda situação. Nosso objetivo referia-se a uma demonstração de que a análise de sentimento poderia trazer benefícios a maneira dos SRIs apresentarem seus resultados. Isto é, elaboramos uma ideia de um instrumento de auxílio à apresentação da lista de resultados de um SRI qualquer por meio da aplicação de uma classificação facetada.

Na seção de demonstração mostramos como seria nossa proposta na prática: identificando as sentenças positivas e negativas em cada documento da lista de resultados dos SRIs, expondo a quantidade delas por documento e realizando a marcação dessas sentenças no próprio documento analisado.

Desta forma, o usuário poderia simplesmente optar pela polaridade que lhe interessasse, e, além disso, ao abrir o documento escolhido, consiguiria enxergar rapidamente as opiniões manifestadas no texto.

Assim sendo, chegamos a conclusão que tal proposta poderia, realmente, auxiliar as pessoas durante suas pesquisas. Porém, é claro, essa ideia ainda é um tanto quanto limitada. Por conta disso, elaboramos uma seção com possíveis trabalhos futuros, aplicações que poderiam incrementar nossa teoria para que se tornasse uma ferramenta mais "poderosa".

REFERÊNCIAS

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[2] D. A. Grossman and O. Frieder, *Information Retrieval: Algorithms and Heuristics*, 2nd ed., ser. The Kluwer International Series of Information Retrieval. Springer, 2004.

[3] F. Costa and L. Ramos, "AnÁlise facetada: Em busca de uma classificaÇÃo para o teatro," *PontodeAcesso*, vol. 2, no. 3, 2008. [Online]. Available: http://www.portalseer.ufba.br/index.php/revistaici/article/view/3215

[4] A. M. D. Tristão, G. R. Fachin, and O. E. Alarcon, "Sistemas de classificação facetados e tesauros: instrumentos para organização do conhecimento," *Ciência da Informação*, vol. 33, no. 2, 2004. [Online]. Available: http://revista.ibict.br/cienciadainformacao/index.php/ciinf/article/view/88/82

[5] W. Zheng and H. Fang, "A retrieval system based on sentiment analysis," in *Proceedings of the Fourth Workshop on Human-Computer Interaction and Information Retrieval*. New Brunswick, NJ, USA: Rutgers University, 2010, pp. 52–54.

[6] V. Broughton, "The need for a faceted classification as the basis of all methods of information retrieval," *Aslib Proceedings*, vol. 58, no. 1/2, pp. 49–72, 2006. [Online]. Available: http://dx.doi.org/10.1108/00012530610648671

[7] J. Mills, "Faceted classification and logical division in information retrieval." *Library Trends*, vol. 52, no. 3, pp. 541–570, 2004. [Online]. Available: http://dblp.uni-trier.de/db/journals/libt/libt52.html#Mills04

[8] R. Prieto-Díaz, "Implementing faceted classification for software reuse," *Commun. ACM*, vol. 34, no. 5, pp. 88–97, May 1991. [Online]. Available: http://doi.acm.org/10.1145/103167.103176

[9] O. Cardoso, "Recuperação de informação," *INFOCOMP Journal of Computer Science*, vol. 2, no. 1, pp. 33–38, 2000.

[10] M. A. Hearst, *Search User Interfaces*, 1st ed. Cambridge University Press, 2009. [Online]. Available: http://searchuserinterfaces.com/book/

[11] A. Barbosa, "Classificações facetadas," *Ciência da Informação*, vol. 1, no. 2, 1972. [Online]. Available: http://revista.ibict.br/ciinf/index.php/ciinf/article/view/1665

[12] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*, 2010.

[13] ——, *Sentiment Analysis and Opinion Mining*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[14] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.

[17] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[15] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," in *Proceedings of the 14th International Conference on World Wide Web*, ser. WWW '05. New York, NY, USA: ACM, 2005, pp. 342–351. [Online]. Available: http://doi.acm.org/10.1145/1060745.1060797

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," Tech. Rep., 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.9562

[18] R. Narayanan, B. Liu, and A. Choudhary, "Sentiment analysis of conditional sentences," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, ser. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 180–189. [Online]. Available: http://dl.acm.org/citation.cfm?id=1699510.1699534

[19] M. Ganapathibhotla and B. Liu, "Mining opinions in comparative sentences," in *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, ser. COLING '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 241–248. [Online]. Available: http://dl.acm.org/citation.cfm?id=1599081.1599112

[20] W. Zhang, L. Jia, C. Yu, and W. Meng, "Improve the effectiveness of the opinion retrieval and opinion polarity classification," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08. New York, NY, USA: ACM, 2008, pp. 1415–1416. [Online]. Available: http://doi.acm.org/10.1145/1458082.1458309

[21] M. Zhang and X. Ye, "A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '08. New York, NY, USA: ACM, 2008, pp. 411–418. [Online]. Available: http://doi.acm.org/10.1145/1390334.1390405

[22] M. Kobayashi and K. Takeda, "Information retrieval on the web," *ACM Comput. Surv.*, vol. 32, no. 2, pp. 144–173, Jun. 2000. [Online]. Available: http://doi.acm.org/10.1145/358923.358934

[23] C. Chen and Y. Yu, "Empirical studies of information visualization: a meta-analysis," *International Journal of Human-Computer Studies*, vol. 53, no. 5, pp. 851–866, 2000.

[24] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing, Second Edition*, N. Indurkhya and F. J. Damerau, Eds. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2010, iSBN 978-1420085921.

[25] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, ser. LSM '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 30–38. [Online]. Available: http://dl.acm.org/citation.cfm?id=2021109.2021114

[26] T. T. Thet, J.-C. Na, and C. S. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," *J. Inf. Sci.*, vol. 36, no. 6, pp. 823–848, Dec. 2010. [Online]. Available: http://dx.doi.org/10.1177/0165551510388123

[27] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010.

[28] V. S. Jagtap and K. Pawar, "Analysis of different approaches to sentence-level sentiment classification." Valletta, Malta: International Journal of Scientific Engineering and Technology, Abril 2013.