

# Uma aplicação de Rede Social de Consumo Baseada em uma Arquitetura de Data Warehouse

Holisson S. da Cunha<sup>1</sup>, Sergio L. S. Mergen<sup>1</sup>

<sup>1</sup> Campus Alegrete - Universidade Federal do Pampa (UNIPAMPA)  
Caixa Postal 97546-550 – Alegrete – RS – Brasil

holissonsud@gmail.com, sergiomergen@unipampa.edu.br

**Abstract.** *This paper describes a data warehousing architecture targeted at the problem of collecting and analyzing products offers available in the Web. Besides presenting the architecture, we describe the modules that were implemented in order to build an social network application that wraps the access to the data warehouse. One of the implemented modules allows collecting data interactively, by gathering information provided by the users of the social network themselves. Another module that deserves attention uses the k-means classifier to automatically divide the users in communities based on their financial profile. Experiments show how the classifier behaves for dividing users in a fixed number of communities given a controlled testing scenario.*

**Resumo.** *Esse artigo descreve uma arquitetura de data warehouse voltada para o problema da coleta e análise de dados de ofertas de produtos disponibilizados na Web. Além de apresentar a arquitetura, serão descritos os módulos que foram implementados tendo em vista a criação de uma aplicação de rede social que encapsula o acesso ao data warehouse. Um dos módulos implementados permite a coleta de dados de maneira interativa, a partir de dados de consumo informados pelos usuários da rede social. Outro módulo que merece destaque utiliza o algoritmo de classificação k-means para automaticamente dividir os usuários em comunidades com base no seu perfil de consumo. Os experimentos relatam como o algoritmo se comporta na divisão dos usuários em um número fixo de comunidades dado um cenário de teste controlado.*

## 1. Introdução

Com a popularização da internet e o avanço das tecnologias de informação, o comércio eletrônico tem crescido com rapidez. Apenas em 2011 foram gastos R\$ 18,7 bilhões em compras pela internet, o que permite visualizar o aumento da procura de serviços e produtos pelos usuários.

A qualidade no serviço e o melhor preço são os principais critérios para a definição de uma compra. As diversas possibilidades que o mercado eletrônico oferece tornam essa tarefa cansativa e demorada, o que desmotiva os usuários.

Devido à dificuldade em definir um local que ofereça melhor preço e maior qualidade de serviço, usuários costumam buscar informações que auxiliem a tomada dessa decisão. Algumas possibilidades são: pesquisas em portais de comparação de preço, sites de *reviews* e redes sociais, onde na maioria das vezes, essas informações são cedidas através da experiência de consumo de outros usuários.

Com base nesse cenário, propõe-se a construção de uma aplicação no formato de uma rede social, que incentiva o usuário a publicar dados pessoais referentes a consumo, aliada a tecnologia de *data warehouse*.

Devido a sua flexibilidade e fácil entendimento, a modelagem multidimensional própria dos *data warehouses* permite agregar dados de forma eficaz, de modo que diversas análises possam ser realizadas [Raghu Ramakrishnan 2003]. No caso em questão, os dados de diferentes usuários a respeito de suas compras podem ser agrupados para a geração de análises estatísticas úteis para os usuários, como gráficos de comparação de preço de produtos.

Neste artigo será apresentada a arquitetura de *data warehouse* proposta. Trata-se de uma arquitetura genérica, que possibilita a coleta de dados de ofertas de produtos e a posterior análise dessas ofertas. No entanto, neste artigo será demonstrada uma aplicação específica, em que a arquitetura serviu para o armazenamento e análise de informações referentes a compras realizadas pelos usuários da rede social.

Dois módulos específicos foram desenvolvidos. Um deles realiza a coleta de dados interativamente, a partir de dados de consumo fornecidos pelos próprios usuários da rede social. O segundo módulo utiliza o algoritmo de classificação não supervisionado *k-means* para agrupar usuários em comunidades que possuam perfis de consumo semelhante. Esse tipo de agrupamento permite aproximar usuários que adquiriram produtos semelhantes, o que pode ser útil na tomada de decisão referente a compra de determinado produto.

Este artigo está dividido da seguinte forma: na seção 2 é apresentada a arquitetura de *data warehouse* proposta, o que inclui a modelagem multidimensional utilizada, o módulo responsável pela carga de dados e o módulo que implementa a técnica *K-means*. Na seção 3 são discutidos experimentos a respeito do uso do *K-means* para o agrupamento de usuários em um número fixo de comunidades. Os trabalhos relacionados são descritos na seção 4. Para finalizar, a seção 5 traz as considerações finais.

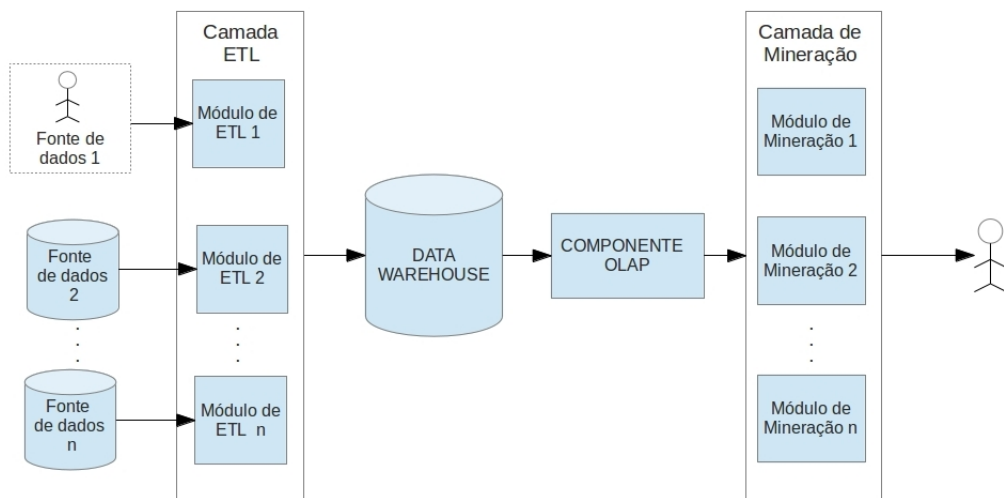
## **2. Arquitetura da Aplicação**

A Figura 1 apresenta a arquitetura proposta para esse trabalho. De modo geral, as ofertas de produtos são extraídos das fontes de dados e alimentados em um *data warehouse*, que decompõe os dados em dimensões que caracterizam cada oferta. A partir dos dados, é possível realizar operações OLAP para gerar gráficos mostrando correlações entre os dados, ou para análises de dados mais sofisticadas através de técnicas de mineração de dados.

Nas próximas seções, os componentes da arquitetura são apresentados de forma detalhada. Juntamente com a explicação dos componentes da arquitetura, serão apresentados os módulos específicos que foram implementados tendo em vista uma aplicação de rede social que encapsula o acesso aos dados armazenados.

### **2.1. Fontes de Dados**

De modo geral, os dados de interesse são aqueles relacionados a ofertas de produtos, o que inclui informações como categoria e preço, por exemplo. Esses dados encontram-se distribuídos em diversas fontes externas, como bases de dados operacionais, planilhas



**Figura 1. Arquitetura de *Data Warehouse* usada para Coleta e Análise de Dados de Ofertas de Produtos**

eletrônicas, documentos XML, formulários Web e sites de *e-commerce*, dentre outros. Neste artigo, destaca-se a possibilidade de recuperar dados de interesse diretamente a partir dos próprios usuários da rede social, conforme ilustrado no retângulo tracejado da Figura 1.

## 2.2. Camada ETL (Extract, Transform and Load )

A carga das fontes de dados para o *data warehouse* é realizada através da camada de ETL. Essa camada é composta por módulos de ETL, sendo que cada módulo se especializa na carga de dados de um tipo específico de fonte de dados.

De modo geral, um módulo de ETL é responsável pelas operações de extração, limpeza e carga. A extração e a carga se encarregam de selecionar os dados da fonte de dados e armazená-la no *data warehouse*, respectivamente. Já a transformação se encarrega do processo de adequação dos dados para que esses possam ser aproveitados. Essa etapa é necessária uma vez que as fontes podem apresentar erros, anomalias e inconsistências. Para solucionar esses problemas, a etapa de transformação utiliza técnicas de limpeza de dados (*data cleaning*), que tem como objetivo detectar e remover anomalias dos dados; e Desambiguidade de dados (*Data Disambiguation*), que busca categorizar corretamente esses dados. Através dessas técnicas é possível alcançar unicidade de informação e obter melhores benefícios de suporte à decisão.

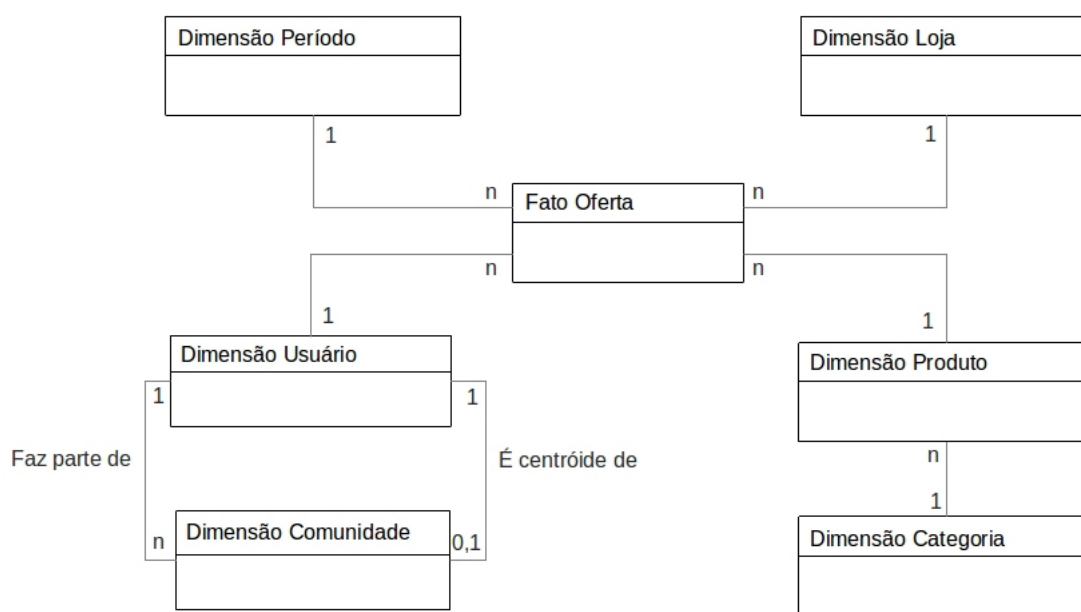
Neste artigo foi projetado um tipo específico de módulo de ETL que recebe dados diretamente de usuários, e não de uma fonte de dados concreta já existente. De acordo com esta abordagem, usuários fornecem suas informações de consumo através de um formulário de registro de compra de produtos, disponível no ambiente de uma rede social.

Para preencher o formulário, os seguintes campos são fornecidos: Produto comprado, Categoria do produto comprado, Período da compra, Loja onde foi feita a compra e Preço do produto comprado. O formulário é disponibilizado como um recurso de um organizador financeiro. Depois de registrar a compra, o usuário pode pesquisar em um calendário o histórico de compras realizadas.

### 2.3. Data Warehouse

O esquema de *data warehouse* utilizado neste trabalho baseia-se no modelo floco de neve, onde as tabelas dimensões principais se relacionam diretamente com a tabela fato, e algumas tabelas dimensão extras se relacionam com as dimensões principais. Essa estrutura tem o propósito de normalizar as tabelas dimensionais, visando reduzir o espaço ocupado por essas tabelas e criar hierarquias entre as mesmas [Machado 2010].

A figura 2 ilustra a modelagem de dados proposta para esse trabalho. O modelo contém uma entidade central (fato), chamada oferta, que armazena os registros de ofertas de consumo fornecidos pelos usuários. A tabela oferta contém o valor unitário de cada produto, e identificadores para as entidades dimensões.



**Figura 2. Modelo Multidimensional Relacionado a Dados de Ofertas de Produtos**

Nas tabelas dimensões são armazenados dados descritivos relacionados às ofertas de produtos, que são o produto ofertado, local onde foi feita a oferta, período da oferta e usuário que adquiriu o produto ofertado. Observe que a dimensão usuário só será preenchida nos casos em que a oferta for na verdade um item já adquirido, sendo que o usuário nesse caso é a pessoa que realizou a compra.

Para fins de normalização, as dimensões categoria e comunidade se relacionam respectivamente com as dimensões produto e usuário. A primeira identifica a categoria do produto ofertado, enquanto a segunda identifica a comunidade a que o usuário pertence. Conforme indicado na figura, existe um segundo relacionamento entre usuário e comunidade, que identifica quem é o usuário central da comunidade. O preenchimento dos dados relacionados a comunidades é visto na seção 2.5.

### 2.4. Componente OLAP

O componente de OLAP (*On-line Analytical Processing*) é responsável pelo acesso aos dados multidimensionais armazenados no *data warehouse*. Diferentemente do modelo

relacional puro, em que as consultas são especificadas através de operadores de seleção, projeção, junção e agrupamento, no modelo multidimensional, onde os dados são dispostos em estruturas dimensionais chamadas de cubos, novos operadores podem ser utilizados. Algumas das operações OLAP mais comuns são *Drill Down*, para diminuir o nível de granularidade de uma dimensão, *Roll up*, para aumentar o nível de granularidade de uma dimensão e *Slice and Dice*, para fatiar o cubo de dados através da seleção em uma das dimensões.

Através das operações OLAP, é possível realizar uma série de análises sobre os dados de consumo fornecidos pelos usuários. Por exemplo, os dados podem ser utilizados para a geração de gráficos de tendência de preço, exibindo as correlações existentes entre locais de venda e período das ofertas. Neste artigo, as operações OLAP são usadas para recuperar os dados que serão utilizados na classificação das comunidades, conforme descrito na próxima seção.

## 2.5. Camada de Mineração de Dados

A camada de mineração de dados utiliza os dados sumarizados armazenados no *data warehouse* para descobrir padrões e correlações entre os fatos automaticamente. Uma das formas de realizar essa análise envolve o uso de algoritmos de aprendizado de máquina, que buscam classificar os fatos em grupos que partilhem de alguma característica em comum.

Neste trabalho, foi implementado um módulo de mineração que divide os usuários em comunidades, agrupando-os de acordo com o seu perfil de consumo. Para isso, foi utilizado o *K-means*, um algoritmo não supervisionado de aprendizado de máquina.

Dado um número de comunidades, o algoritmo mapeia os usuários em um espaço multidimensional e executa um número limitado de iterações. Em cada iteração são escolhidos os usuários centróides de cada comunidade. Os demais usuários são agrupados na comunidade onde a distância euclidiana entre ele e o centróide for menor.

Na primeira iteração, os centróides são escolhidos aleatoriamente. Nas iterações seguintes, o centróide de uma comunidade será o usuário central da comunidade calculada na iteração anterior. O agrupamento termina após um número pré-definido de iterações ou depois que a classificação convergir, ou seja, quando os agrupamentos não modificarem entre uma iteração e a seguinte.

Para o cálculo da distância euclidiana, os usuários são mapeados em um espaço multidimensional composto pelas categorias de produtos existentes no *data warehouse*. Para cada usuário, o vetor de categorias  $\{c_1, c_2, \dots, c_n\}$  é preenchido de acordo com as compras feitas pelo usuário. Ou seja, o valor de  $c_i$  é definido como sendo a soma de todos os produtos comprados que sejam da categoria  $c_i$ . A seleção dos fatos, categorias e usuários é feita através dos operadores de OLAP mencionados na seção anterior.

Após o término da execução do algoritmo, o *data warehouse* é atualizado com a atribuição da comunidade de cada usuário, assim como com a atribuição do usuário centróide de cada comunidade, conforme os dois relacionamentos entre as dimensões usuário e comunidade, indicados na Figura 2.

## 2.6. Visualização das Comunidades

O resultado da divisão em comunidades pode ser visualizado através de um recurso fornecido pela rede social. Ao acessar a visualização, um usuário obtém uma lista ordenada de comunidades, sendo que a ordem é relativa a distância euclidiana entre o usuário e os centróides das comunidades. A primeira comunidade será aquela a que o usuário efetivamente pertence. As demais comunidades são aquelas que possuem graus decrescentes de relevância com o usuário. Essa forma de exibição leva em consideração que o perfil de um usuário pode se espalhar por diversos grupos de interesse, e que é interessante que ele tenha acesso à pessoas que pertençam a todas as comunidades definidas.

Como a classificação é não supervisionada, os usuários são divididos em grupos que partilhem características em comum. No entanto, ainda é necessário identificar cada uma dessas comunidades com uma descrição breve e precisa. A forma encontrada de realizar essa tarefa foi através de um gerador de assinaturas. Esse gerador cria uma *tag cloud* (nuvem etiquetada) para as comunidades, contendo o nome de produtos. Os produtos que foram mais comprados pelos membros das comunidades estarão com o tamanho de fonte maior e cor diferenciada, permitindo que o usuário identifique o propósito de cada comunidade.

Através da Figura 3, pode-se visualizar como as comunidades serão apresentadas para usuário. Cada membro de uma comunidade será representado através da imagem do seu avatar, seguido do seu nome definido no cadastro da rede social. É possível ver que o usuário está mais fortemente relacionado a comunidade em que o produto *Netbook* é mais largamente vendido.

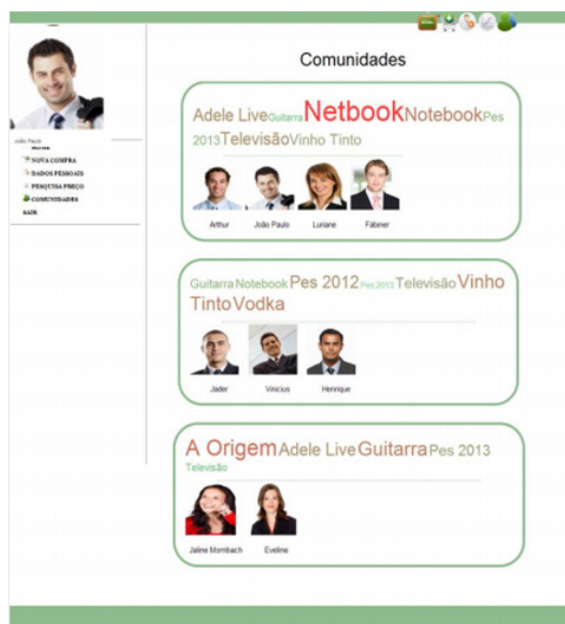


Figura 3. Comunidades geradas com base no perfil de consumo dos usuários.

## 3. Experimentos

Esta seção apresenta os resultados da avaliação do nível de precisão para a classificação de usuários com base no seu perfil de consumo, utilizando o algoritmo de aprendizado de

máquina *K-means*. Neste trabalho, apresentaremos os resultados obtidos utilizando um número de centróides fixo equivalente a três. Ou seja, os usuários serão agrupados em três comunidades distintas ao final da etapa de classificação.

Para realizar o experimento foram armazenados no *data warehouse* registros fictícios de compras. De modo a comparar o resultado da classificação de forma objetiva, foi usado um cenário de teste controlado, de acordo com os critérios definidos abaixo:

- Foi gerado um conjunto de usuários  $U_1 \leftarrow \{u_1, u_2, u_3, u_4\}$ .
- Foi gerado um conjunto de usuários  $U_2 \leftarrow \{u_5, u_6, u_7, u_8\}$ .
- Foi gerado um conjunto de usuários  $U_3 \leftarrow \{u_9, \dots, u_{20}\}$ .
- Foi gerado um conjunto de categorias  $C_1 \leftarrow \{c_1, c_2, c_3, c_4\}$ .
- Foi gerado um conjunto de categorias  $C_2 \leftarrow \{c_5, c_6, c_7, c_8\}$ .
- Foi gerado um conjunto de categorias  $C_3 \leftarrow \{c_9, \dots, c_{20}\}$ .
- Para cada usuário  $u_i \in U_1$ , foram gerados oito registros de compras, sendo duas compras para cada categoria de  $C_1$
- Para cada usuário  $u_i \in U_2$ , foram gerados oito registros de compras, sendo duas compras para cada categoria de  $C_2$
- Para cada usuário  $u_i \in U_3$ , foram gerados oito registros de compras da categoria  $c_i$

Os critérios acima foram escolhidos de modo que cada um dos conjuntos de usuários apresente comportamento distinto dos demais. Os usuários de  $U_1$  realizaram compras de produtos de quatro categorias, enquanto os usuários de  $U_2$  realizaram compras de produtos de outras quatro categorias. Já os usuários de  $U_3$  são aqueles que compram produtos de categorias que ninguém mais adquiriu, ou seja, o perfil de consumo de cada um desses usuários é distinto de todos os demais.

A Figura 4 apresenta a classificação que foi obtida para essa configuração. Como pode-se ver, o classificador reconheceu que os usuários dos conjuntos  $U_1$  e  $U_2$  possuem perfis de consumo semelhantes, e agrupou todos os demais usuários em um *cluster* a parte. Idealmente esses usuários não deveriam ser considerados próximos de ninguém, dado que seus perfis de consumo são divergentes. No entanto, o uso de apenas três centróides impede que isso aconteça.

Mesmo assim, é interessante ver que o classificador pelo menos não agrupou esses usuários com aqueles que possuem perfil bem definido. Dado que o número de centróides é fixo, pode-se considerar esse resultado satisfatório, uma vez que o terceiro grupo pode ser interpretado como O resultado gerado pode ser interpretado como o conjunto de pessoas que compram produtos de apenas uma categoria.

#### 4. Trabalhos Relacionados

No contexto deste trabalho, técnicas de classificação aplicadas sobre dados de produtos e baseadas no algoritmo *K-means* já vem sendo aplicadas para problemas de segmentação de mercado. Em alguns casos, o algoritmo é usado em conjunto com outras técnicas, como mapas auto organizáveis [Kuo et al. 2002] e acasalamento de abelhas (*Honey Bee Mating*) [Fathian and Amiri 2008]. No problema de segmentação de mercado, a partir da análise do comportamento, os clientes são separados em segmentos, que são posteriormente marcados como rentáveis ou não. Nesse caso, o objetivo da classificação é basicamente direcionar campanhas de marketing para segmentos mais rentáveis. Já o nosso

		CATEGORIAS																				
		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	
USUÁRIOS	U1	2	2	2	2																	
	U2	2	2	2	2																	
	U3	2	2	2	2																	
	U4	2	2	2	2																	
	U5					2	2	2	2													
	U6					2	2	2	2													
	U7					2	2	2	2													
	U8					2	2	2	2													
	U9									8												
	U10										8											
	U11											8										
	U12												8									
	U13													8								
	U14														8							
	U15															8						
	U16																8					
	U17																	8				
	U18																		8			
	U19																			8		
	U20																				8	

**Figura 4. Análise da Classificação Realizada pelo *k-Means***

trabalho tem a própria aproximação das pessoas como objetivo. Essa distinção faz com que as técnicas partam de informações diferentes para realizar a classificação, como o dia da compra, o número de compras e o valor dos produtos adquiridos. Consequentemente, a própria concepção de que pessoas devem compor uma comunidade tende a diferir.

O trabalho em questão também pode ser analisado como um problema de recomendação, onde o objetivo é indicar pessoas com quem indivíduos específicos gostariam de se relacionar. A literatura está repleta de estudos relacionados a problemas de recomendação baseado no histórico de ações do usuário. Essa área recebe o nome de Filtragem Colaborativa (*Collaborative Filtering*). Em boa parte dos casos, o objetivo envolve a recomendação de produtos, como no site de compras da Amazon [Linden et al. 2003]. A partir das preferências conhecidas de alguns usuários, pretende-se fazer recomendações com base em preferências desconhecidas de outros usuários.

Algumas técnicas de recomendação, chamadas de técnicas baseadas em modelos, usam algoritmos de classificação como o *K-means* para realizar a predição de produtos [Su and Khoshgoftaar 2009]. No entanto, conforme descrito em [Linden et al. 2003], o processo de recomendação *online* (em tempo real) se torna oneroso quando embasada em algoritmos de aprendizado de máquina.

Também existem trabalhos de Filtragem Colaborativa que visam a recomendação de pessoas. Para esses casos, normalmente a recomendação pode ser feita em modo *offline*, o que permite o uso de técnicas menos responsivas, como as baseadas em classificação. As abordagens propostas variam de acordo com o objetivo específico que se pretende atingir. Por exemplo, em [Ungar et al. 1998], pessoas são agrupadas com base em seus interesses por filmes. Como a informação de gosto é esparsa, foi proposta uma forma de classificação que ocorre em duas frentes. Numa delas são agrupados filmes que são preferidos pelo mesmo grupo de pessoas, e na outra são agrupadas pessoas que preferem os mesmos filmes.

Existem também trabalhos de recomendação que atuam diretamente sobre re-



des sociais. Em [Hannon et al. 2010], são gerados relacionamentos entre usuários do *Twitter* com base na análise do conteúdo dos seus Tweets. Já em [Cai et al. 2010], a recomendação de pessoas usa a rede de relacionamentos existentes, explorando o conceito de gosto (de quem uma pessoa gosta) e atratividade (quem gosta de uma pessoa).

Convém destacar que o uso de *data warehouse* com dados coletados a partir da Web já foi assunto de investigação na literatura. Em [Hernández et al. 2011], esse modelo é chamado de *data webhouse*. Conforme descrito no artigo, os dados coletados são oriundos da navegação do usuário dentro de um Web site corporativo, e o *data warehouse* gerado tem como objetivo principal a tomada de decisões estratégicas que visam aprimorar a experiência da navegação, de modo que o usuário encontre a informação que procura mais rapidamente.

O modelo que propomos neste artigo também pode ser definido como um *data webhouse*, uma vez que os dados podem ser coletados diretamente através da interação do usuário com a rede social. No entanto, os dados armazenados podem ser utilizados para um número maior de aplicações, como aquelas voltadas para a segmentação de mercado ou recomendação de pessoas/produtos, conforme descrito nos trabalhos acima apresentados.

## 5. Conclusão

Esse trabalho apresentou uma arquitetura de *data warehouse* para coleta e análise de ofertas de produtos. Para validar a arquitetura, foi desenvolvida uma aplicação que utiliza serviços de uma rede social tanto para a coleta como para a análise de dados. Especificamente, a rede social dá ênfase a dados de consumo registrados pelos seus usuários.

Para demonstrar as análises que os dados armazenados possibilitam, o algoritmo *K-means* foi empregado para realizar a separação dos usuários em comunidades, levando em consideração seu histórico de compras registradas na rede social. Essa separação pode ser usada de diversas formas, como por usuários interessados em produtos específicos e por empresas interessadas em campanhas publicitárias focadas. Por exemplo, compras coletivas podem partir tanto do conjunto de pessoas interessadas em um produto quanto por uma empresa que perceba esse interesse nas comunidades que monitora.

A aplicação de compra coletiva é apenas um exemplo que pode ser explorado. O que vale a pena destacar é que, para que aplicações como essas sejam possíveis, é necessário ter os dados concentrados de uma forma que facilite a análise posterior. Nesse sentido, o artigo propõe uma abordagem prática ao empregar redes sociais tanto para alimentar o *data warehouse* quanto para consumir os dados após análise.

Como trabalhos futuros, pretende-se incorporar novos módulos de ETL à arquitetura, de modo que mais informações de ofertas estejam disponíveis. Os módulos podem ser ativos, acessando fontes de dados na Web em busca das informações, ou passivos, sendo chamados por agentes externos que realizam a carga. Essa opção é interessante para lojas que desejem ter seus produtos a disposição na aplicação.

Além disso, pretende-se aprimorar o módulo de ETL existente, para que os dados dos usuários sejam consolidados antes de alimentarem o *data warehouse*. Um dos métodos a ser estudado envolve usar técnicas de casamento de string para corrigir o nome de um produto, caso outro nome semelhante seja encontrado no *data warehouse*.

Outra possibilidade de trabalho futuro envolve modificar o algoritmo *K-means* utilizado de modo que o número de comunidades seja dinâmico. Conforme apresentado nos experimentos, o uso de um número fixo de comunidades pode agrupar usuários que não possuem nenhuma relação de consumo aparente. Também deseja-se estudar as dimensões que compreendem o *data warehouse* afim de identificar novos tipos de análises que podem ser feitas sobre os dados, o que inclui estender o algoritmo de classificação de usuários usando outras dimensões além das categorias de produtos que foram adquiridos.

## Referências

- Cai, X., Bain, M., Krzywicki, A., Wobcke, W., Kim, Y. S., Compton, P., and Mahidadia, A. (2010). Learning collaborative filtering and its application to people to people recommendation in social networks. In *Proceedings of the 2010 IEEE International Conference on Data Mining, ICDM '10*, pages 743–748, Washington, DC, USA. IEEE Computer Society.
- Fathian, M. and Amiri, B. (2008). A honeybee-mating approach for cluster analysis.
- Hannon, J., Bennett, M., and Smyth, B. (2010). Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems, RecSys '10*, pages 199–206, New York, NY, USA. ACM.
- Hernández, P., Glorio, O., Garrigós, I., and Mazón, J.-N. (2011). Towards a model-driven framework for web usage warehouse development. In *Proceedings of the 30th international conference on Advances in conceptual modeling: recent developments and new directions, ER'11*, pages 336–337, Berlin, Heidelberg. Springer-Verlag.
- Kuo, R. J., Ho, L. M., and Hu, C. M. (2002). Integration of self-organizing feature map and k-means algorithm for market segmentation. *Comput. Oper. Res.*, 29(11):1475–1493.
- Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.
- Machado, F. N. R. (2010). *Tecnologia e Projeto Data Warehouse*. Sao Paulo-SP, 3 edition.
- Raghu Ramakrishnan, J. G. (2003). *Sistemas de Gerenciamento de Banco de Dados*. Sao Paulo-SP, 3 edition.
- Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2.
- Ungar, L., Foster, D., Andre, E., Wars, S., Wars, F. S., Wars, D. S., and Whispers, J. H. (1998). Clustering methods for collaborative filtering. AAAI Press.