5-2015

Multimedia Content Recommendation in Digital Convergence Environments: An Approach Based on Data Mining and Semantic Web

Priscilla Kelly Machado Vieira
Universidade Federal Rural de Pernambuco (UFRPE), priscillakmv@gmail.com

Natasha Correia Queiroz Lino
Universidade Federal da Paraíba (UFPB), natasha@ci.ufpb.br

Recomendação de Conteúdos Multimídia em Ambientes de Convergência Digital: Uma Abordagem Baseada em Mineração de Dados e Web Semântica

Alternative Title: Multimedia Content Recommendation in Digital Convergence Environments: An Approach Based on Data Mining and Semantic Web

Priscilla Kelly M. Vieira
Universidade Federal Rural de Pernambuco (UFRPE)
Unidade Acadêmica de Garanhuns
Garanhuns, PE, Brasil
priscillakmv@gmail.com

Natasha Queiroz Lino
Universidade Federal da Paraíba (UFPB)
Centro de Informática
João Pessoa, PB, Brasil
natasha@ci.ufpb.br

RESUMO

Com o advento da TV Digital interativa (TVDi), nota-se o aumento de interatividade no processo de comunicação além do incremento das produções audiovisuais, elevando o número de canais e recursos disponíveis para o usuário. Esta realidade faz da tarefa de encontrar o conteúdo desejado uma ação onerosa e possivelmente ineficaz. A incorporação de sistemas de recomendação no ambiente TVDi emerge como uma possível solução para este problema. Este trabalho tem como objetivo propor uma abordagem híbrida para recomendação de conteúdo em TVDi, baseada em técnicas de Mineração de Dados, integradas a conceitos da Web Semântica, permitindo a estruturação e padronização dos dados e consequente possibilidade do compartilhamento de informações, provendo semântica e raciocínio automático. Para o serviço proposto é considerado o Sistema Brasileiro de TV Digital (SBTVD) e o *middleware* Ginga. Foi desenvolvido um protótipo e realizado experimentos com a base de dados do NetFlix, utilizando a métrica de precisão para avaliação. Obteve-se uma precisão média de 30%, utilizando apenas a técnica de mineração. Acoplando-se com as regras semânticas obteve-se precisão média de 35%.

Palavras-Chave

Sistemas de Recomendação, Mineração de Dados, TVDi, Representação de Conhecimento.

ABSTRACT

The emerging scenario of interactive Digital TV (iDTV) is promoting the increase of interactivity in the communication process and also in audiovisual production, thus raising the number of channels and resources available to the user. This reality makes the task of finding the desired content becoming a costly and possibly ineffective action. The incorporation of recommender systems in the iDTV environment is emerging as a possible solution to this problem. This work aims to propose a hybrid approach to content recommendation in iDTV, based on data mining techniques, integrated to the semantic web concepts, allowing structuring and standardization of data and consequently making possible sharing of information, providing semantics and automated reasoning. For the proposed service it is considered the Brazilian Digital TV System (SBTVD) and the middleware Ginga. A prototype has been developed and experiments carried out with a NetFlix database. As results, it was obtained an average accuracy of 30% using only the data mining technique. On the other hand, the evaluation including semantic rules obtained an average accuracy of 35%.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: On-line Information Services – data sharing, web-based services

General Terms

Algorithms, Design.

Keywords

Recommender Systems, *Data Mining,* iDTV, Knowledge Representation.

1. INTRODUÇÃO

Com o advento da TV Digital interativa (TVDi), nota-se o aumento de interatividade no processo de comunicação além do incremento das produções audiovisuais [1], o que torna o ambiente favorável ao desenvolvimento de aplicações dedicadas a esta tecnologia. Neste sentido, a TVDi promoveu e possibilitou um cenário de aumento da quantidade de canais, serviços e conteúdo disponíveis ao usuário. Esta nova realidade faz com que a tarefa de encontrar o conteúdo desejado se torne uma ação onerosa e, em alguns casos, ineficiente. É neste contexto que Sistemas de Recomendação [2; 3; 4] emergem como solução possível para auxiliar esta escolha.

Neste sentindo, é proposto neste trabalho um serviço de recomendação de conteúdo no ambiente do Sistema Brasileiro de TV Digital (SBTVD) [5], baseado em Mineração de Dados e em conceitos e técnicas da Web Semântica. Estas duas estratégias foram escolhidas por já serem utilizadas para o propósito deste

trabalho, mas não terem sido avaliadas em conjunto, considerando o ambiente brasileiro [6; 7]. Para a esquematização do serviço de recomendação de conteúdo foi definida e instanciada uma arquitetura, tal como projetada uma ontologia responsável por estruturar dados multimídia minerados.

É importante destacar que as estratégias e arquitetura utilizadas neste trabalho, podem ser aplicados e avaliadas em diversos outros ambientes com o objetivo de recomendação de conteúdo. Contudo, foi utilizado ambiente de TVDi como forma de demonstração e validação das estratégias propostas.

O restante deste artigo está organizado da seguinte forma: a seção 2 apresenta uma fundamentação teórica de Sistemas de Recomendação de conteúdo. Na seção 3 é explanada a arquitetura do serviço de recomendação, o *Recommender Knowledge TV*, tal como o detalhamento de cada um dos componentes envolvidos. Na seção 4 são destacados os experimentos e discussões. Em seguida, na seção 5, são explanados os trabalhos relacionados. Por fim, a seção 6 traz as conclusões e as perspectivas de trabalhos futuros.

2. SISTEMAS PARA RECOMENDAÇÃO DE CONTEÚDO

Sistemas de Recomendação são comumente aplicados em sistemas web [8], mas nos últimos anos começou-se o estudo de recomendação de conteúdo no ambiente da TVDi por possuir algumas peculiaridades.

A plataforma de TVDi tem um conjunto de requisitos que a torna desafiadora para a construção de sistemas de recomendação de conteúdo. Primeiramente, o ambiente de TV pode se mono-usuário ou multi-usuário, por isso ao recomendar é difícil identificar o tipo de experiência que está sendo realizada e, consequentemente, identificar o usuário que solicita o serviço. Em segundo lugar, este é um cenário multi-plataforma, onde o conteúdo pode ser entregue a diferentes plataformas computacionais, tais como, set-top-boxes (STBs) com o *middleware* padrão, dispositivos móveis e de computação pessoal em geral. Adicionalmente, o comportamento do usuário ao acessar conteúdos destinados a TVDi é, em geral, passivo, diferindo do comportamento ativo na web. Neste sentindo, o serviço de recomendação proposto neste trabalho está de acordo com as características intrínsecas ao ambiente da TVDi.

É importante destacar que Sistemas de Recomendação para a TVDi pode possuir diferentes perspectivas de uso, dentre elas destaca-se: (i) do usuário; para sistemas que objetivam auxiliar o usuário na busca por conteúdos de seu interesse. (ii) do provedor de conteúdo; para auxiliar na construção de grades de programação para maximizar alguma métrica, por exemplo audiência. (iii) de marketing, para dar suporte a veiculação de comerciais para o público alvo correto. Este trabalho foca na perspectiva do usuário.

Adicionalmente, é considerado um ambiente de convergência digital. Uso conjunto de duas plataformas distintas: TVDi e web, fazendo uso de dados da web para compor uma solução na TVDi. Na próxima seção são explanadas algumas características do processo de recomendação de conteúdo proposto neste trabalho.

3. PROCESSO PROPOSTO

Para este trabalho foi considerado o SBTVD [5], incluindo o *middleware* Ginga [9]; camada de software intermediária, presente no receptor, e situada entre as aplicações e o sistema operacional, que oferece uma série de facilidades para o desenvolvimento de conteúdo e aplicativos para TV Digital Interativa [10; 11]. Neste sentido propõe-se uma arquitetura para recomendação de conteúdo em ambientes de convergência (TVDi e web), considerando as duas plataformas em conjunto, a TVDi conectada.

No âmbito deste trabalho é utilizado um processo híbrido de recomendação de conteúdo [12] que consiste no uso conjunto de Filtragem Colaborativa (FC) e Filtragem Baseada em Conteúdo (FBC).

Na estratégia com FC é realizada a troca de experiências entre usuários. São detectados grupos de usuários com comportamentos semelhantes, e a recomendação de conteúdo é realizada entre estes, o que possibilita recomendações inesperadas e novas ao usuário. A FBC parte do princípio de que, se um usuário aprovou um determinado programa, é provável que no futuro goste de outros similares, com as mesmas características [12].

A abordagem híbrida foi escolhida por 3 (três) motivos: (i) validar o objetivo específico de permitir recomendação multimídia em TVDi conectada (ambiente de convergência digital), considerando, assim, os dois cenários definidos na literatura [13], FC e FBC. (ii) Validar a arquitetura genérica de recomendação de conteúdo proposta neste trabalho, garantindo que esta pode ser instanciada para ambos os tipos de recomendação de conteúdo. (iii) Considerar o desafio de recomendação de conteúdo multimídia no ambiente da TVDi, como sendo um ambiente de experiência multiusuário, permitindo a atenuação dos problemas das abordagens em separado.

Considerando tais características explanadas, neste trabalho a tarefa de clusterização é realizada sobre os dados dos históricos dos usuários (FC). Neste contexto, supondo que o usuário y (Uy) solicita uma recomendação, é detectado qual o usuário x (Ux) do cluster possui maior similaridade com o Uy. Um conteúdo visto por Ux é recomendado para Uy.

Como forma de aperfeiçoar a decisão de qual programa recomendar para um Uy, a partir de um Ux, regras semânticas são utilizadas (seção 3.2) e as recomendações são realizadas entre os usuários do mesmo cluster (FBC). As regras detectam qual programa de Ux é mais similar semanticamente com os programas vistos pelo Uy.

Com o objetivo de possibilitar o processo de recomendação descrito, foi especificada e avaliada uma arquitetura para prover o serviço de recomendação de conteúdo, o *Recommender Knowledge TV* (RKTV), introduzido na próxima seção.

3.1 Arquitetura

Sob a perspectiva da execução das operações previstas neste trabalho, tais como a tarefa de Mineração de Dados, considerando o ambiente da TVDi conectada, o *middleware* Ginga e a capacidade de processamento dos decodificadores de sinais, os STBs, este trabalho propõe uma arquitetura cliente-servidor, na qual o módulo cliente coleta dados do usuário,

situados no STB (STB) do usuário e inserido dentro do núcleo comum do *middleware*, e o módulo servidor é responsável pela maior parte do processamento de informações e a disponibilização do serviço. Esta arquitetura foi assim definida por ter sido analisado o funcionamento de aplicações na plataforma da TVDi brasileira, onde não se tem o controle da sobreposição de dados e a seguridade de que o contexto do usuário é armazenado ao ser alterado o canal. Uma forma de garantir a manutenção do contexto do usuário é inserindo esta funcionalidade no Núcleo Comum do Ginga [14].

Na Figura 1 são apresentados os componentes de cada um dos módulos da arquitetura RKTV (*Recommender Knowledge TV*).

O módulo cliente coleta dados por meio de dois componentes (Figura 1): (i) *Monitor Agent*: Monitora o comportamento dos usuários do STB em relação ao consumo do conteúdo exibido na TVDi conectada. (ii) *Provider Agent*: Tem a função de capturar as informações sobre a programação visualizada pelo o usuário.

O *Monitor Agent* [14], para monitorar o comportamento dos usuários do STB, está em comunicação constante com o componente *Tuner* [9] do núcleo do *middleware* Ginga (*Ginga Common Core*), responsável pela sintonização dos canais.

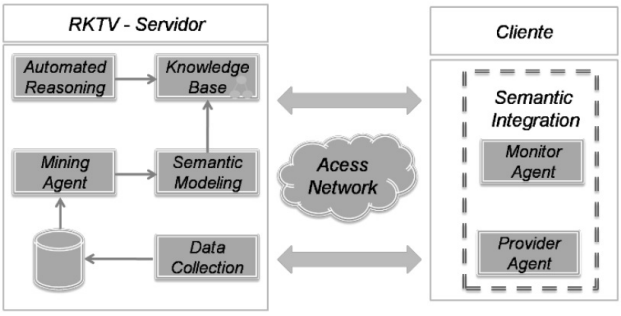

Figura 1. Arquitetura Cliente-servidor do RKTV.

O módulo *Provider Agent* [14] recebe as interações do usuário por meio do módulo *Monitor Agent* e captura os metadados da programação via web ou componente SI do Ginga Common Core [9], responsável por extrair os metadados transmitidos pelos provedores de conteúdos.

Para o objetivo principal deste trabalho alguns dados mínimos devem ser coletados a partir do *Provider Agent*. Estes devem ser utilizados nos demais módulos especificados, auxiliando o serviço de recomendação de conteúdo. Para isto, foram adaptados os metadados definidos em [15], reduzindo seu número e adequando-os ao contexto deste trabalho. Na Tabela 1 são listados os metadados a serem obtidas e suas respectivas funções.

O módulo servidor é composto por cinco módulos:

- *Data Collection*: Captura os dados da programação das emissoras e dos usuários (a partir dos componentes inseridos e localizados no núcleo do Ginga Common Core);
- *Mining Agent*: Responsável por todo o processo de Mineração de Dados sobre os dados coletados do STB e da web sobre o usuário e provedores de conteúdo;
- *Semantic Modeling*: Torna homogenia a descrição do conteúdo multimídia minerado, estruturado por meio de ontologias;
- *Knowledge Base*: Armazena todo o conhecimento gerado no processo proposto em forma de instâncias de ontologias;
- *Automated Reasoning*: Responsável por inferir as informações presentes nas bases de conhecimento, gerando recomendações de conteúdo compatíveis aos interesses do usuário.

O Modulo *Data Collection* tem como função principal capturar dados: conteúdo visualizado pelo usuário. Estes dados são baseados nos metadados transmitidos junto ao conteúdo audiovisual, por meio do componente *Service Information* (SI) do *middlaware* Ginga [14], tal como da web. Os dados provenientes do histórico do usuário são enviados pelo componente *Semantic Integration* (Figura 1), via canal de retorno.

Tabela 1. Metadados Extraídos a partir do *Monitor Agent*

Dado	Função
STB_id	Identificar o STB
Program_name	Identificar nome do programa
Channel	Identificar o canal da TV
Program_genre	Gênero do programa
Start_time	Data e horário em que o usuário começou a assistir ao programa
End_time	Data e horário em que o usuário terminou de assistir ao programa

Os metadados capturados a partir do módulo *Data Collection* (Listados na Tabela 1) são depositados em um banco de dados que é submetido ao processo de Mineração de Dados. O Módulo *Mining Agent* é responsável por aplicar a tarefa de Mineração de Dados, tal como a de pré-processamento. Neste trabalho instanciamos apenas a tarefa de clusterização [16] de dados, mas qualquer tarefa de mineração pode ser instanciada nesta arquitetura. Na próxima subseção será detalhada a metodologia utilizada para a criação da ontologia para estruturar conteúdos multimídia minerados, permitindo o compartilhamento de informações tal como raciocínio semântico.

3.2 Representação Semântica

A Web Semântica pode ser definida como uma extensão da web sintática (web simplificada, sem semântica para os dados), na qual é dada à informação um significado [17], permitindo: (i) processamento automático por máquinas, por meio da padronização da representação do conhecimento, (ii) integração de dados, (iii) reuso de dados e (iv) inferência de conhecimento.

A proposta da Web Semântica é baseada nos conceitos de Representação do Conhecimento (RC), subárea da Inteligência Artificial que avalia como o conhecimento pode ser representado simbolicamente e manipulado de forma automática por máquinas [18].

Formalizar o conhecimento permite a interoperabilidade de dados, assim como o raciocínio automático por máquinas, podendo agregar semântica a dados brutos. Para esta formalização, pode-se fazer uso de ontologias [19; 20].

Neste trabalho é utilizada a abordagem da Mineração de Dados para as ontologias, onde o conhecimento proveniente do processo da tarefa de clusterização é representado por meio de uma ontologia, descrita em OWL, chamada OntoRKTV (Onto –

Ontologia e RKTV – *Recommender Knowledge TV*), possibilitando assim, o raciocínio automático sobre os dados e o compartilhamento de dados.

Neste sentido, pode-se realizar raciocínio, por exemplo, sobre gêneros de programas, recomendando, portanto, programas com gênero equivalente ao que o usuário costuma visualizar ou programas com gênero em um subdomínio que o usuário costuma visualizar.

Dentre a gama de metodologias propostas para a construção de ontologias [21; 22], neste trabalho utilizou-se a metodologia 101 especificada por Noy e McGuiness [23] por ser uma metodologia já estabelecida e ter sua implementação e documentação baseada na ferramenta Protegé [24], amplamente difundida para a construção de ontologias.

Seguindo os passos definidos em [23] foi gerada a ontologia da Figura 2, com destaque para o nome das classes e os relacionamentos. As propriedades foram suprimidas para facilitar a visualização geral da OntoRKTV.

Dentre as classes modeladas tem-se: (i) Cluster: Agrupamento de STB com uso semelhante. (ii) STB: Um membro que pode ser pertencente a um cluster. (iii) History: Os membros dos clusters possuem históricos de uso; (iv)PeriodOfDay: Os membros de um cluster são definidos pelo padrão de uso por turno.

As classes ContentDescription e TemporalUnit foram definidas em uma outra ontologia [15] que descreve todas os conceitos importantes para descrever conteúdos multimídia. As classes descritas são utilizadas como links entre os conceitos de clusterização e de dados multimídia para recomendação, indicando, assim, o reuso de ontologias para compor a ontologia descrita neste trabalho. Neste sentido, ContentDescription, ou descrição de conteúdo, é a classe que representa um ponto chave agregador de diversas informações descritivas acerca do conteúdo transmitido. A classe TemporalUnit detalha questões de tempo sobre a programação.

Algumas regras semânticas foram definidas e desenvolvidas sobre a OntoRKTV, para apoiar o processo de recomendação no ambiente da TVDi:

I. *isPartOfCluster*: dois STBs U_1 e U_2 pertencem a um mesmo cluster se: (i) U_1 visualiza conteúdo semelhante ao conteúdo visualizado por U_2.

II. *neighbor*: U_1 é vizinho com forte interseção com U_2 se: (i) O vizinho com maior número de conteúdos similares à U_2 é U_1; e (ii) U_1 pertence ao mesmo cluster que U_2.

A regra *isPartOfCluste*r é uma consequência da tarefa de Mineração de Dados. Dentro de cada cluster, podem-se utilizar as regras que se baseiam diretamente nos relacionamentos oriundos da definição de ontologias:

III. *isEquivalentTo*: uma instância D_1 é equivalente a uma instância D_2 se (i) D_1 é equivalente a um conceito k na ontologia OntoRKTV; (ii) D_2 é equivalente a um conceito w na mesma ontologia; e (iii) k e w estão ligados por um relacionamento equivalentClass na OntoRKTV.

IV. *isSubConceptOf*: um dado D_1 é uma especialização de um dado D_2 se (i) D_1 é idêntico a um conceito k na ontologia OntoRKTV; (ii) D_2 é idêntico a um conceito w na mesma ontologia; e (iii) k e w estão ligados por um relacionamento subClassOf na ontologia OntoRKTV.

V. *isSuperConceptOf*: um dado D_1 é uma generalização de um dado D_2 se: (i) D_1 é idêntico a um conceito k na ontologia OntoRKTV; (ii) D_2 é idêntico a um conceito w na mesma ontologia; e (iii) k e w estão ligados por um relacionamento *superClassOf*.

VI. *isInstanceCloseTo*: Dois dados D_1 e D_2 são instâncias próximas se: (i) D_1 é idêntico a uma instância k na ontologia OntoRKTV e D_2 é idêntico a uma instância w na mesma ontologia; e (ii) k e w são instâncias do mesmo conceito. No contexto deste trabalho, instância é o valor atribuído a uma classe ou propriedade de uma ontologia.

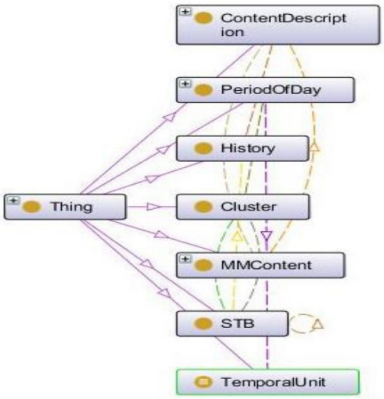

Figura 2. Ontologia OntoRKTV.

As regras definidas em III –VI são utilizadas para avaliação semântica do conteúdo a ser recomendado de um usuário para outro. Neste sentido, avalia-se se dois conteúdos possuem proximidades semânticas. Por exemplo, se um filme é de um subgênero do gênero que um usuário costuma visualizar.

Na próxima subseção será detalhado o módulo *Mining Agent* da arquitetura RKTV.

3.3 Mining Agent

O módulo *Mining Agent* (Figura 1) é responsável por todo o processo relacionado com a descoberta de conhecimento, com foco principal em pré-processamento e Mineração de Dados [25].

Tarefas de Mineração de Dados [26] classificam o padrão que se deseja obter da base de dados. Podem ser classificadas como: (i) Regras de Associação, (ii) Clusterização, (iii) Sumarização, (iv) Classificação e (v) Regressão, dentre outros. Neste trabalho damos foco à tarefa de Clusterização [26; 27], com o objetivo de avaliar o comportamento da recomendação no ambiente de convergência em análise, dado que um outro trabalho [28] avalia este comportamento por meio da tarefa de Regras de Associação. Neste sentido não é objetivo deste trabalho propor algoritmos ótimos de clusterização, mas mostrar a viabilidade da aplicação desta tarefa no ambiente da TVDi.

Adicionalmente, clusterização é utilizada como parte do processo de recomendação proposto por ser amplamente utilizada em sistemas de recomendação e ser considerada de boa precisão para o contexto de recomendação [29].

Existem diversos algoritmos na literatura com foco em clusterização de dados [30; 31]. Neste trabalho é dada ênfase à técnica k-means [30; 31; 32]. Este foi escolhido por ser um algoritmo popular, simples, direto e com custo linear [30].

Para realizar a clusterização de instâncias de objetos é necessário o uso de funções que calculam a similaridade entre os objetos: No contexto de clusterização existem diversas funções que objetivam calcular a distância entre dois objetos [33], dentre estas: euclidiana, *levenshtein*, *hamming*, cosseno. Para a escolha da função é necessário avaliar os dados com relação as características, tipos de dados que serão comparados e os objetivos da comparação [33].

Neste sentido, foi avaliado o comportamento da função de distância euclidiana. Esta foi utilizada nos experimentos explanados na próxima seção (seção 4).

A distância euclidiana é expressa segundo a Equação 1. Nesta, x e y são vetores com n atributos numéricos. Assim (xa – ya) é a diferença entre os atributos de x e y na posição a. A distância euclidiana é a raiz quadrada da soma do quadrado da diferença (xa – ya), para todo a.

Considerando o contexto deste trabalho, os vetores, x e y, que são parâmetros para a função da Equação 1, representam o histórico do usuário x e do usuário y. Na seção 4.1 é destacado o processo para a geração destes vetores. Na próxima seção são detalhados os experimentos para a validação deste trabalho.

Equação 1. Distância euclidiana

$$E(x,y) = \sqrt{\sum_{a=1}^{n} (x_a - y_a)^2}$$

4. Experimentos

A arquitetura RKTV (Figura 1) é genérica, podendo ser instanciada por diversas técnicas e com o uso de diferentes tecnologias. Como forma de demonstrar a viabilidade da arquitetura, foi realizada uma implementação de cada um dos módulos especificados no componente servidor. Neste sentido, foi desenvolvido um protótipo passível de recomendação de conteúdo.

O foco deste trabalho não foi a especificação/implementação do componente do núcleo comum do Ginga (Figura 1). Este foi especificado e implementado em outro trabalho [14], tal qual as necessidades do RKTV, definidas neste trabalho, bem como do projeto mais abrangente do qual o RKTV faz parte.

O módulo *Mining Agent* foi desenvolvido em Matlab, desde o pré-processamento dos dados até a fase da tarefa de Mineração de Dados (onde foi utilizado o algoritmo k-means para clusterização).

As funcionalidades do módulo *Semantic Modeling* foram desenvolvidas por meio da ferramenta Protegé [24]. Para raciocínio automático utilizou-se a API Jena [34].

Por fim, foi desenvolvido em Java um protótipo baseado no RKTV, instanciando e integrando os seus módulos, implementando o processo de recomendação proposto neste trabalho.

Com o objetivo de validar a aplicabilidade do processo definido para o RKTV da plataforma KTV para TVDi, foram realizados experimentos para dar subsídios que, de forma prática, demonstrassem a aplicabilidade e viabilidade deste trabalho. Nesse sentido foi utilizada uma base de dados do serviço de stream de vídeos NetFlix [35] com o registro de histórico de usuários. A utilização da base de dados disponibilizada pelo Netflix se dá devido esta fazer parte de um subdomínio da TVDi, no caso, filmes. Estas características agregam valor a esta pesquisa e experimentos feitos, principalmente por ser uma base de dados real.

A base original de dados do Netflix disponibiliza as seguintes informações: IDFilme (identificação do filme), CustomerID (identificação do usuário), Rating (Número de 'estrelas' atribuídos a um filme por um cliente), Título do filme, Ano que um filme foi lançado e Data de uma classificação.

Os experimentos foram utilizados com a base original (com os atributos listados), e uma repetição do experimento com a base alterada por meio de um enriquecimento semântico. Este enriquecimento se deu por informações retiradas do DBPedia [36], por meio de consultas em SPARQL [37]. Os atributos coletados foram: (I) Diretor, nome do diretor responsável pelo filme. (ii) Ator, atores que participaram do filme; (ii) Categoria, categoria na qual o filme se enquadra.

4.1 Pré–processamento dos dados

No processo de descoberta de conhecimento, uma das etapas que antecede a Mineração de Dados é o pré-processamento. Este precisa ser utilizado para que os dados sejam preparados de acordo com a técnica a ser utilizada. Para a realização da clusterização de usuários, foi gerada uma matriz com uma coluna para o código do usuário e colunas para os filmes vistos por cada usuário. A Tabela 2 demonstra esta transformação. Cada linha da tabela representa um vetor de interesses de um usuário, sendo assim, dois usuários com interesses semelhantes possuem vetores próximos. Desta forma, na Tabela 2, tem-se que o usuário de identificação “69867”, assistiu ao filme “Congo”, uma vez, e aos filmes “Mississippi Burning" e "The Santa Clause", zero vezes.

Tabela 2. Exemplo de vetores representativos dos interesses dos usuários

user_id	“Congo”	“Mississippi burning”	“the santa clause”
69867	1	0	0
437	1	1	0

Com o objetivo de realizar recomendações, a base foi fatiada por períodos, por meio do atributo “Data de Classificação”. Isto foi realizado para que fosse possível realizar recomendações futuras, podendo confirmar quais delas o usuário realmente aceitaria. Neste caso, foram realizados 30 fatiamentos aleatórios, de seis meses cada.

Considerando uma das fatias, com os 6 primeiros meses de uso do sistema ("08-12-1999” - "08-05-2000"). Após o pré-processamento temporal, os dados foram sumarizados da seguinte forma: 23136 Avaliações, 3432 Usuários, 187 Filmes, 168 Categorias, 168 Atores e 155 Diretores.

Analisando os dados, observou-se a ocorrência de *sparcity problem* [8], bastante conhecido em abordagens colaborativas, que caracteriza-se pela dispersão dos itens avaliados pelo o indivíduo em relação a todos os itens do sistema. Este problema ocorre em situações onde o número de itens é muito grande

tornando impraticável as suas avaliações. Este problema se deu pela grande variedade de filmes a serem avaliados, gerando uma matriz esparsa de dados (muitas dimensões com poucas avalições).

Para comprovar esta hipótese o algoritmo k-means foi executado com todas as dimensões geradas, com k, número de clusters, variando de 2 – 400 e o máximo de média de *Silhouette Coefficient* (Coeficiente de Silhueta) [27], utilizado para avaliação de clusters, foi 0.05, quando o indicado é superior a 0,5 [25, 38].

Portanto, foi realizada a redução da variabilidade de filmes, selecionando apenas aqueles que foram visualizados por pelo menos 10% dos usuários do sistema. Como exemplo de redução de uma das execuções do experimento tem-se: 18 dimensões de filmes, 17 de atores, 17 de diretores e 17 categorias.

Este pré-processamento foi realizado em todas as 30 partições avaliadas no processo de validação deste trabalho. Na próxima subseção será avaliado o processo de recomendação proposto neste trabalho.

4.2 Execução do Processo Proposto do *Mining Agent*

Após o pré-processamento dos dados foram aplicadas as técnicas para a detecção do agrupamento de dados. O algoritmo foi executado considerando o histórico de cada usuário.

Para o processamento do algoritmo k-means, o k variou de 2 – 400. Para diminuir a distorção causada pela obtenção de mínimos locais, influência da escolha dos centroides iniciais, [27], foram realizados testes com 5 (cinco) sementes aleatórias diferentes para cada valor de k. Após este processo, selecionou-se a semente que trouxe melhor qualidade aos clusters. Esta qualidade foi avaliada de acordo com duas funções: (i) Média do coeficiente de silhueta [27], que avalia combinação entre coesão e separação dos grupos, (ii) Média dos erros quadráticos [27], que avalia a coesão dos agrupamentos gerados. A média do coeficiente de silhueta foi avaliado segundo [27, 38].

A partir de k = 157, as alterações na média do coeficiente de silhueta foram pouco significativas, com crescimento lento, como demostrado no gráfico da Figura 5. Neste sentido, os experimentos foram realizados com k=157.

Além do coeficiente de silhueta, foi avaliada a medida de erro quadrático, que tende a zero, quando k tende a infinito [26]. É importante destacar que a decisão do número de grupos é primordial. Não se deve quebrar um cluster em dois distintos, nem tão pouco unir clusters que essencialmente são distintos [16]. Desta forma, avalia-se o erro quadrático onde ocorre um pequeno "joelho" e uma diminuição na velocidade que o erro decrementa [27], como destacado no gráfico da Figura 6.

É importante destacar, que neste trabalho foram avaliadas duas medidas para seleção do k ideal, no entanto, estas são independentes, convergindo para um mesmo resultado aproximado, k = 157. Contudo, outros valores de k, aquém de 157, podem ser igualmente escolhidos, no entanto, observou-se um ganho baixo, dado que o número de clusters unitários torna-se bastante presente e pouco representativa para o contexto de recomendação.

Na execução do experimento com a base de dados enriquecida, obteve-se resultados semelhantes, não houveram melhoras significativas com relação ao coeficiente de silhueta, tão pouco uma redução na soma dos erros quadráticos. Isto se deu devido a características dos dados enriquecidos, pois poucos filmes possuem mesmo ator, diretor ou categoria. O enriquecimento realizado por meio do DBPedia, foi bastante específico, o que não influenciou positivamente no agrupamento dos dados. Esta característica do enriquecimento resultou em uma matriz esparsa, mesmo depois do pré-processamento explanado na subseção 4.1. Neste sentido, os dados provenientes do enriquecimento foram melhor utilizados com as regras semânticas, visto na seção 4.3.

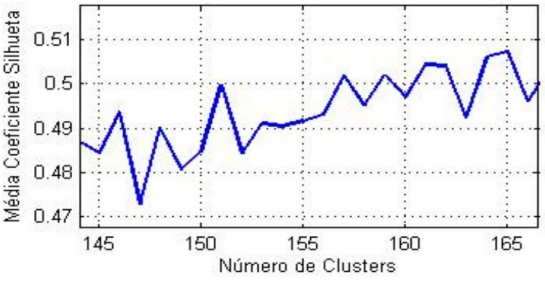

Figura 5. Gráfico Média Coeficiente Silhueta.

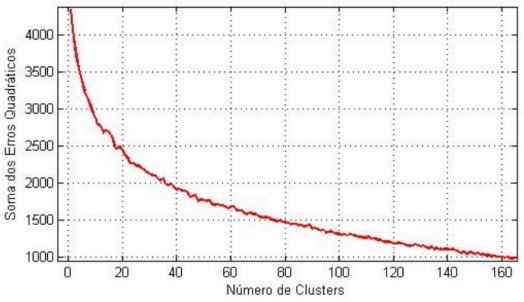

Figura 6. Gráfico Soma dos Erros Quadráticos

4.3 Uso das Regras Semânticas

Após o processo de clusterização (agrupamento sintático), foi integrado ao processo de recomendação do conteúdo o uso das regras semânticas definidas e formalizadas na seção 3.2. Estas regras foram implementadas e avaliadas neste trabalho.

No contexto da base de dados avaliada neste trabalho, as regras definidas em I - VI (seção 3.2) foram aplicadas sobre os atributos que representam a categoria dos filmes, atores e diretores. Considerando os dados da Tabela 3, por exemplo, não ocorre especialização/generalização de gêneros, mas ocorre equivalência de diretores entre os filmes "Yojimbo" e "Seven Samurai" e entre "Presumed Innocent" e "The Devil's Own".

No contexto deste trabalho as regras semânticas foram utilizadas com o objetivo de reduzir o grupo de filmes a serem recomendados, tal como aumentar a precisão do sistema, como descrito e avaliado na próxima subseção.

4.4 Análise

Para analisar a recomendação do processo proposto neste trabalho, foi utilizada a métrica *Precision* (Precisão) [39], medindo a porcentagem de itens recomendados de forma correta ao usuário. No contexto deste trabalho, foi realizada uma média da precisão da recomendação realizada aos usuários que foram selecionados aleatoriamente durante os experimentos.

Neste sentido, a precisão das recomendações para cada usuário, segue a expressão definida na Equação 2. A Precisão (P) por usuário é a divisão entre as o número de recomendações que o sistema acertou ($R_{Acertadas}$), pelo número de recomendações realizadas pelo processo (R_{Total}).

Tabela 3. Exemplo do resultado de regras semânticas aplicadas durante os experimentos

Nome do filme	Categoria	Diretor
Yojimbo	Jidaigeki	Akira Kurosawa
Seven Samurai	Fictional samurai	Akira Kurosawa
Presumed Innocent	Works by Scott Turow	Alan J. Pakula
The Devil's Own	Troubles	Alan J. Pakula

Equação 2. Expressão de precisão por usuário

$$P = \frac{R_{Acertados}}{R_{Total}}$$

A precisão total do processo proposto está definida na expressão da Equação 3. Precisão total do processo proposto neste trabalho (PT) é a média aritmética das precisões de cada usuário, tal que n é o número de usuários.

Equação 3. Expressão da precisão total

$$P_t = \frac{\sum_{i=1}^{i=n} \frac{R_{Acertados\,i}}{R_{Totais\,i}}}{n}$$

Os acertos foram avaliados a partir do histórico de cada usuário, e das informações provenientes da própria base do Netflix. O processo descrito na subseção anterior foi replicado 30 vezes, com períodos diferentes. Em média, obteve-se uma precisão de 30,04%, sem considerar o uso das regras semânticas.

Adicionando a análise semântica, como previsto no processo de recomendação de conteúdo proposto neste trabalho, obteve-se um refinamento do conjunto de filmes recomendados pelo RKTV, atingindo, em média, 35,12% na métrica de precisão. Este refinamento se dá pela diminuição do conjunto de elementos a serem recomendados. É importante destacar que o pequeno ganho de precisão da recomendação apenas com a clusterização e da recomendação seguida da avaliação semântica se deu pelas características do enriquecimento semântico dos dados. Estes possuem poucas interseções de categorias, tal como poucos mapeamentos exatos definidos nas instâncias na OntoRKTV. Na próxima seção são destacados trabalhos que possuem similaridades com a pesquisa explanada neste artigo.

5. TRABALHOS RELACIONADOS

Aroyo [6] propõe a incorporação da semântica na TVDi, considerando-a no ambiente da web, a fim de possibilitar o desenvolvimento de aplicações distribuídas e a personalização do serviço. Aroyo baseia seu trabalho de personalização nos conceitos da Web Semântica e de Representação do Conhecimento. Assim como Aroyo, nossa proposta também é baseada nos conceitos da Web Semântica, no entanto, propomos recomendação no ambiente do SBTVD, incorporando no *middleware* Ginga um módulo que dá suporte ao processo de recomendação proposto.

Ávila [7] apresenta um sistema integrado ao *middleware* Ginga baseado em mineração por regras de associação, entretanto sem nenhuma estruturação semântica do conteúdo. Diferentemente, este trabalho propõe o uso de ontologias para estruturar o conhecimento após o processo de clusterização dos dados.

Neto[40] e Kim [29] também propõem recomendação de conteúdo utilizando Mineração de Dados. O sistema de recomendação proposto por Kim, Recommendation System of IPTV TV Program Using Ontology and K-means Clustering, apresenta uma solução baseada em mineração de dados por clusterização, mas neste caso a mineração ocorre sobre ontologias.

6. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou uma proposta de abordagem baseada em conhecimento e Mineração de Dados para recomendação de conteúdo em TV Digital interativa conectada. Utilizando o *middleware* Ginga como estudo de caso, propôs-se a especificação de um processo de recomendação que integra conceitos e técnicas da Web Semântica e Mineração de Dados para filtragem híbrida de conteúdo. É importante destacar que inúmeras abordagens e técnicas de recomendação de conteúdo podem ser expressas por meio da arquitetura do RKTV. Neste trabalho, é iniciada a investigação sobre o comportamento de técnicas de clusterização e semântica em conjunto.

Como trabalho futuro, é importante destacar a necessidade de avaliação de outras tarefas de Mineração de Dados, tal como a extensão da OntoRKTV. Adicionalmente, destaca-se a necessidade de especificação do raciocínio automático a ser realizado sobre a ontologia. Por fim, ainda são necessários experimentos para ratificar a importância da OntoRKTV no processo de recomendação. Adicionalmente, é possível incorporar ao processo de recomendação deste trabalho a análise sobre a audiência dos conteúdos, como por exemplo em [41].

7. REFERÊNCIAS

[1] Médola, A. S. L.(2009) Televisão Digital Brasileira e os Novos Processos de Produção de Conteúdos: Os Desafios para o Comunicador. Revista da Associação Nacional dos Programas de Pós-Graduação em Comunicação | E-compós, Brasília, v.12, n.3, set./dez.

[2] Resnick, P. and Varian, H. R. (1997) Recommender systems. Communications of the ACM. ACM 40, 3 (Mar. 1997), 56-58.

[3] Xu, M.; Berkovsky, S.; Ardon, S.; Triukose, S.; Mahanti, A.; Koprinska, I. Catch-up TV Recommendations: Show Old Favourites and Find New Ones. RecSys '13. p. 285-294.

[4] Azaria, A.; Hassidim, A.; Kraus, S.; Eshkol, A.; Weintraub, O.; Netanely, I. Movie Recommender System for Profit Maximization. RecSys '13. p. 12-128.

[5] SBTVD – Sistema Brasileiro de TV Digital. Ministério das Comunicações. Disponível em: <http://sbtvd.cpqd.com.br/>. Acessado em Janeiro de 2014.

[6] Aroyo, L.; Bellekens, P.; Björkman, M.; Houben, G.; Akkermans, P.; Kaptein, A.; (2007) SenSee Framework for Personalized Access to TV Content. European Conference on Interactive TV (EuroITV 2007), 156-165.

[7] Ávila, P. M. (2010) RecommenderTV: Suporte ao Desenvolvimento de Aplicações de Recomendação para o Sistema Brasileiro de TV Digital. Dissertaçao de Mestrado. Universidade Federal de São Carlos, Brasil.

[8] Torres, R. Personalização na Internet. Editora Novatec, 2004.

[9] Souza Filho, G. L. de, Leite, L. E. C., Batista, C. E. C. F. Ginga-J: The Procedural Middleware for the Brazilian Digital TV System. In: Journal of the Brazilian Computer Society, no 4, Vol 12, (ISSN 0104-6500) pp. 47-56. Março, 2007. Porto Alegre, RS, Brasil.

[10] Soares, L. F.; Lemos, G. (2007) Interactive Television in Brazil: System Software and the Digital Divide. In European Interactive TV Conference - EuroITV2007. Amsterdam.

[11] Lekakos, G.; Chorianopoulos, K.; Doukidis, G. Interactive Digital Television: Technologies and Applications. IGI Publishing, 2007.

[12] Chorianopoulos, K.. Personalized and mobile digital TV applications. Proc. 2008 Multimedia Tools and Applications, Volume 36, Issue 1-2. Kluwer Academic Publishers, pp. 1-10, 2008.

[13] Adomavicius, G.; Tuzhilin, A. Towards the Next Generation of Recommender Systems: A survey of the State-of-the-Art and Possible Extensions. Knowledge and Data Engineering, IEEE Transactions, Volume 17, Issue 6, page(s): 734– 749, June 2005.

[14] Silva, C. F.; Vieira, P. K. M.; Lino, N. C. Q. Semantic Integration - Uma Extensão do Núcleo do Middleware Ginga. SBSI, 2013.

[15] Araújo, J. P. C. (2011) CoreKTV - Uma infraestrutura baseada em conhecimento para TV Digital Interativa: um estudo de caso para o middleware Ginga. Dissertação de Mestrado. Universidade Federal da Paraíba.

[16] Berkhin. P.. Survey of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[17] Berners-Lee, T.; Lassila, O.; Hendler, J. The semantic web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 2001.

[18] Russell, S. and Norvig, P. Artificial Intelligence - A Modern Approach. 2 edição, Prentice Hall, 2002.

[19] Breitman, K.. Web Semântica: O Futuro da Internet. 1. ed. Rio de Janeiro: LTC - Livros Tecnicos e Cientificos Editora S.A., 2005. v. 1. 190 p.

[20] Guarino, N. (1995) Formal Ontology, Conceptual Analysis and Knowledge Representation, International Journal of Human-Computer Studies, 43(5-6):625–640.

[21] Fernández-Lópes, M.; Gómez-Pérez, A. Overview and analysis of methodologies for building ontologies. In: Journal The knowledge Engineering Review. Vol 17, pp 129-156. New York, 2002.

[22] Gómez-Pérez, A. Evaluation of taxonomic knowledge in ontologies and knowledge bases. In Proc. Of the 12th Workshop on Knowledge Acquisition, Modeling and Management, KAW'99. Voyager Inn, Banff, Alberta, Canada. 1999.

[23] Noy, F. N.; McGuinness, D. L. Ontology Development 101: A Guide to Creating Your First Ontology'. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

[24] Protegé. Disponível em: <http://protege.stanford.edu/>. Acessado em Junho de 2014.

[25] Kononenko, I. & Kukar, M. Machine learning and data mining. Chichester, UK: Horwood Publishing, 2007.

[26] Han, J. and Kamber, M. Data Mining Concepts and Techniques. 2a Edição, Editora Elsevier, Reino Unido. 2006.

[27] Tan, P.N., Steinbach, M., Kumar, V. Introduction to Data Mining, Addison-Wesley, 2006.

[28] Patrício Júnior, J. C. A. Mining Knowledge TV: Uma Abordagem de Ambiente de KDD com Ênfase em Mineração de Dados no Ambiente da Knowledge TV. Dissertação de Mestrado. Universidade Federal as Paraíba, 2012.

[29] Kim, J., Kwon, E., Cho, Y., Kang, S. Recommendation System of IPTV TV Program Using Ontology and K-means Clustering. Second International Conference, UCMA 2011, Daejeon, Korea, April 13-15, 2011. Proceedings, Part II.

[30] Berry, M.J.A.; Linoff, G. Data mining techniques. John Wiley & Sons, Inc. 1997.

[31] Ochi, L. S., Dias, C. R., Stênio, S. F.. Clusterização em Mineração de Dados. Escola Regional de Informática Rio de Janeiro. Espírito Santo. Mini Curso. Rio das Ostras, 2004.

[32] Jain, A. K., Murty, M. N., and Flynn, P. J.. Data clustering: a review. ACM Comput. Surv, 1999. 31(3):264–323.

[33] Wang, H.; Nearest Neighbor by Neighborhood Counting. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.28, no.6, pp. 942-953. 2006.

[34] Jena. Disponível em: < http://jena.sourceforge.net/>. Acessado em Janeiro de 2014.

[35] NetFlix. Disponível em <www.netflix.com>. Acessado em Junho de 2014.

[36] BDPedia. Disponível em <http://pt.dbpedia.org/>. Acessado em: Janeiro de 2014.

[37] DuCharme, B. Learning SPARQL. O'Reilly Media, 2011.

[38] Khalilian, M., Mustapha, N., Suliman, N. , Mamat, A. A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets. IMECS, 2010. Hong Kong.

[39] Montaner, M. Collaborative Recommender Agents Based on Case-Based Reasoning and Trust. 2003.

[40] Neto, M. M., Cardoso, D., Teixeira, C. T., Cortés, M. Abordagem Combinada para Recomendação Personalizada Utilizando o Guia de Programação Eletrônico. CSBC' 2010, Belo Horizonte, MG, 2010.

[41] Basilio, A. C A.; Moreno, M. F.; Barrere, E.Análise de Interação e Audiência em Sistemas de TV Digital Interativa. Webmedia, 2012.