Um Modelo de Recomendação de Arquivos para Sistemas de Armazenamento em Nuvem

Alternative title: A File Recommendation Model For Cloud Storage Systems

Ricardo Batista Rodrigues
Centro de Informática
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brasil
rbr@cin.ufpe.br

Carlo M. R. da Silva
Centro de Informática
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brasil
cmrs@cin.ufpe.br

Frederico A. Durão
Universidade federal da Bahia (UFBA)
Salvador, BA, Brasil
fred@gmail.com

Rodrigo E. Assad
Universidade Federal Rural de Pernambuco (UFRPE)
Recife, PE, Brasil
assad@usto.re

Vinicius C. Garcia
Centro de Informática
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brasil
vcg@cin.ufpe.br

Silvio R. L. Meira
Centro de Informática
Universidade Federal de Pernambuco (UFPE)
Recife, PE, Brasil
srlm@cin.ufpe.br

RESUMO

O desenvolvimento tecnológico vivenciado nos últimos anos proporcionou o crescimento do universo digital de forma exponencial, parte desse universo digital encontra-se armazenado em sistemas de armazenamento em nuvem. A cada dia surgem mais destes sistemas, que oferecem o armazenamento de dados de forma distribuída com alta taxa de disponibilidade, o que tem impucionado cada vez mais usuários a migrarem seus dados para a nuvem. No entanto, a grande quantidade de arquivos armazenada nestes sistemas dificulta a filtragem de conteúdo relevante, demandando tempo e trabalho por parte do usuário na busca por arquivos com conteúdo similar as suas preferências. Diante deste cenário, esta pesquisa propõe um modelo de recomendação para sistemas de armazenamento em nuvem, que tem como objetivo utilizar características da nuvem associadas à técnica de recomendação baseada em conteúdo.

Palavras-Chave

Recomendação, armazenamento em nuvem, computação em nuvem.

ABSTRACT

The technological development in recent years has experienced the exponentially growth of the digital universe, part of this digital universe lies stored in cloud storage systems. With each day, more of these systems come out, offering data storage in a distributed manner with the proposal to provide high availability rate, what has driven more and more users who have migrated your data to the cloud. However, the large amount of files stored in these systems makes it difficult to filter relevant content, requiring time and labor by the user in searching for files with similar content to your preferences. Face of this scenario, this study proposes a model for recommendation of files in cloud storage systems, which aims to use cloud features associated with the technique of content-based recommendation.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Systems; H.3 [**Information Storage and Retrieval**]: Miscellaneous

General Terms

Theory

Keywords

Recommendation, cloud storage, cloud computing.

1. INTRODUÇÃO

Vivemos em uma era de efervescência informacional, a cada dia se produz mais informação e, geralmente, estas informações são armazenadas em meios digitais. O tamanho do universo digital cresce de forma exponencial. Segundo relatório publicado pela EMC Corporation[1] [22], em 2005, o volume de dados chegou a 130 exabytes; em 2010, superou 1 zettabyte e a previsão é que em 2015 chegue a quase 8 zettabytes [16].

Este universo digital citado em Gantz e Reinse [16] expande e torna cada vez mais complexa a tarefa de filtragem de conteúdo relevante que atenda às preferências do usuário. Dentre as técnicas utilizadas na filtragem de conteúdo podemos citar os sistemas de recomendação[26].

[1]http://brazil.emc.com/index.htm?fromGlobalSelector

Quando é preciso filtrar um grande conjunto de dados, podemos utilizar técnicas de recomendação para facilitar o processo de filtragem de informações relevantes e similares com as preferências do usuário. Para isto, é necessário informações sobre o indivíduo alvo da recomendação ou sobre o ambiente que influenciará na geração da recomendação. A partir daí, um sistema de recomendação poderá recomendar os itens que mais se aproximam das preferências ou características do usuário ou do seu ambiente [1] [19].

A partir da análise da literatura e dos sistemas de armazenamento em nuvem mais populares e utilizados atualmente, é notório que esses sistemas não fornecem ao usuário o serviço de recomendação de arquivos. Na maioria dos sistemas de armazenamento em nuvem, a filtragem de conteúdo é realizada por sistemas de busca, onde o usuário fornece termos chaves e o sistema retorna arquivos com o título ou conteúdo similar aos termos apresentados pelo usuário.

Por outro lado, a enorme quantidade de sistemas de recomendação que funcionam como serviço em nuvem, não utilizam características da nuvem na geração de suas recomendações. Esses sistemas normalmente possuem como objetivo recomendar itens que atendam as preferências dos usuários, sem considerar requisitos do ambiente. Neste trabalho investiga-se a utilização de características da nuvem que podem ser usadas no processo de recomendação de arquivos em ambiente de armazenamento em nuvem.

2. FUNDAMENTAÇÃO TEÓRICA

O National Institute of Standards and Technology (NIST)[2] define computação em nuvem como um modelo que permite que um conjunto de recursos computacionais possam ser fornecidos sob demanda de forma a permitir que os mesmos sejam fornecidos e liberados rapidamente com o mínimo de esforço de gestão ou interação do fornecedor [11] [14] [23].

Entre os recursos disponiveis na técnologia em nuvem, está o de armazenamento, o qual provê recursos e serviços de armazenamento baseados em servidores remotos que utilizam os princípios da computação em nuvem [32]. Armazenamento em nuvem tem duas características básicas: a primeira trata da infraestrutura da nuvem, a qual baseia-se em clusters de servidores baratos; a segunda tem o objetivo de, através dos *clusters* de servidores, armazenamento distribuído e redundância de dados, fazer múltiplas cópias dos dados armazenados para alcançar dois requisitos: alta escalabilidade e alta usabilidade [9]. A alta escalabilidade significa que o armazenamento em nuvem pode ser dimensionado para um grande aglomerado com centenas de nós ou peers de processamento. Alta usabilidade significa que o armazenamento em nuvem pode tolerar falhas de nós e que estas falhas não afetam todo o sistema [12].

2.1 Sistemas de Recomendação

De acordo com Lima at al. [2], Sistemas de Recomendação são ferramentas e técnicas de software que fornecem recomendação de itens [26]. As sugestões geradas por um SR podem estar relacionadas a processo de tomada de decisão em diversos contextos, como a escolha de itens em *e-commerces*, conteúdo similar as preferência do usuário, livros, amigos em redes sociais, rotas geograficas e etc [2] [28].

Uma recomendação pode se basear nas preferências de quem a faz e pode ser dirigida a um indivíduo específico, ou para um público maior. Para a pessoa que recebe a recomendação, ela funciona como um filtro ou uma visão particular de um universo de possibilidades geralmente inacessível. Ela pode levar em consideração também a preferência de quem está à procura de sugestões e não apenas de quem a faz. É possível até mesmo fazer recomendação baseada nas opiniões de outras pessoas. Alguém que não é admirador do gênero Rock pode recomendar discos baseado no que seus amigos que apreciem tal estilo costumam ouvir. Ainda, a recomendação pode incluir explicações sobre como ela foi gerada para permitir que o seu recebedor a avalie [25] [26].

No universo dos sistemas de recomendação, existem diversas técnicas e formas de gerar recomendação, dentre quais destacam-se quatro técnicas como principais e mais utilizadas: recomendação baseada em conteúdo, filtragem colaborativa e filtragem híbrida [2] [6] [15] [26].

Filtragem baseada em conteúdo: esta categoria de sistemas recomenda ao usuário itens semelhantes àqueles em que ele demonstrou interesse no passado. Para tanto, o sistema analisa as descrições dos conteúdos dos itens avaliados pelo usuário para montar o seu perfil, o qual é utilizado para filtrar os demais itens da base. Esse conteúdo no qual ele se baseia são elementos explícitos como nome, descrição, *tags*, conteúdo, categorização ou *rating* do item a ser recomendado. Os resultados são o julgamento da relevância daqueles itens para o usuário, e a consequente recomendação ou não [24] [26]. Uma das vantagens de implantar o Filtro Baseado em Conteúdo é que a quantidade de usuários no sistema não interfere na eficácia do SR, já que se baseia somente no histórico do que o usuário já acessou. Em contrapartida, um sistema assim precisa de itens bem descritos, com informação suficiente para categorizá-los. Outro problema encontrado nesse tipo de recomendação é a sugestão de itens sempre muito parecidos, limitando os usuários de conhecer itens diferentes [26] [25].

Filtragem colaborativa: a abordagem de recomendação por filtragem colaborativa foi proposta inicialmente para suprir as deficiências da abordagem baseada em conteúdo. Com o passar dos anos, conquistou tamanha aceitação que hoje é provavelmente a técnica mais amplamente conhecida, implementada e utilizada para sistemas de recomendação [3] [17]. Na abordagem colaborativa, em contraste com a recomendação baseada em conteúdo, a compreensão ou conhecimento do conteúdo dos itens é totalmente prescindível. Ao invés de buscar itens disponíveis com conteúdos similares aos previamente avaliados positivamente pelo usuário para indicá-los, ela se apoia inteiramente na similaridade entre os usuários do sistema para o processo de sugestão. Partindo do princípio de que as melhores recomendações para um indivíduo são aquelas feitas por pessoas com preferências similares às dele, o sistema identifica estas pessoas para sugerir itens que as mesmas tenham aprovado e ainda não tenham sido consumidos pelo indivíduo [26] [27].

Filtragem híbrida: são sistemas de recomendação que utilizam duas ou mais técnicas, para amenizar os problemas apresentados por cada técnica [26] [27].

3. TRABALHOS RELACIONADOS

Existem alguns trabalhos na literatura que discutem e apresentam sistemas de recomendação em nuvem. Nesta seção, serão apresentados alguns SRs destacando o modelo de recomendação utilizado, objetivando avaliar as contribuições e diferenciais desta pesquisa.

[2]http://www.nist.gov/srm/

Lai et al. [20] que apresentam um sistema de recomendação de programas de televisão (TV) baseado em computação em nuvem e um *framework map-reduce*. Esta proposta de arquitetura tem como objetivo ofertar um *backend* escalável para suportar a demanda de processamento de dados em larga escala para um sistema de recomendação. No que tange os usuários, Lai et al. [20] os agrupam de acordo com suas preferências, cada programa de TV recebe um peso que é atribuído de acordo com o período de tempo que o usuário o assistiu. A popularidade de um programa é indicada pelo seu peso, os programas populares em um grupo de usuários, são recomendados para usuários de outros grupos que tenham semelhanças de preferências entre si. Nesta pesquisa os autores propõe a utilização de técnicas de computação em nuvem para lidar com grandes conjuntos de dados, devido ao seu poder computacional e de estrutura escalável.

Jung et al. [19], apresentam a plataforma CloudAdvisor de recomendação em nuvem. A proposta desta plataforma é recomendar configurações de nuvem de acordo com as preferências do usuário como orçamento, expectativa de desempenho e economia de energia para determinada carga de trabalho. Permitindo ainda, que o usuário faça comparação das recomendações recebidas, como qual é o melhor preço para a carga de serviço desejada. A plataforma tem como objetivo auxiliar o usuário na escolha dos melhores serviços e proporcionar aos provedores de serviço em nuvem à oportunidade de adequação as expectativas e preferências dos usuários.

Existem diversos sistemas de recomendação disponíveis na Internet, e boa parte destes sistemas estão relacionados à nuvem, seja como parte de sistemas em nuvem ou hospedados em servidores em nuvem. Muitos deste utilizam dados da nuvem em suas recomendações ou características da nuvem para gerar recomendações. A proposta desta pesquisa se diferência das demais descritas nesta seção por utilizar características da nuvem na geração de suas recomendações. Desta forma, cada critério é parte da recomendação, que tem como objetivo proporcionar aos usuários a melhor utilização dos recursos em nuvem disponíveis.

4. O MODELO DE RECOMENDAÇÃO

O modelo de recomendação proposto neste trabalho, é composto por cinco critérios, que foram utilizados no processo de recomendação. Os critérios propostos foram definidos a partir da observação de sistemas de armazenamento em nuvem. Os critérios são: Similaridade, Disponibilidade, Taxa de *Download*, Tamanho do Arquivo e Popularidade do Arquivo.

Critério Similaridade: este critério atende ao requisito referente às preferências do usuário. Neste critério, é calculado a similaridade entre o conteúdo de um arquivo no qual o usuário tenha demonstrado preferência e arquivos armazenados em nuvem, que são candidatos a serem recomendados. Para calcular a similaridade entre os conteúdos dos arquivos, é proposto a utilização da técnica de similaridade do cosseno, que retorna um valor entre 0 (zero) e 1 (um). Esta abordagem, foi proposta por ser utilizada com frequência na avaliação de semelhança entre dois itens [4] [8] [21]. O cálculo de similaridade do cosseno é apresentado pela Equação 1:

$$St = cos(\theta) = \frac{A \cdot B}{\parallel A \parallel \cdot \parallel B \parallel} \quad (1)$$

Na Equação 1, é calculada a similaridade entre dois arquivos, o conteúdo de cada um é representando por um vetor (vetores A e B), de onde se obtém o produto de "A" e "B" e calcula-se a magnitude dos dois vetores. As magnitudes, são multiplicadas e divididas pelo produto escalar dos vetores "A" e "B". Os arquivos que sejam similares ao arquivo que represente as preferências do usuário serão ranqueados de acordo com o seu grau de similaridade, ou seja, quanto maior o *score* de similaridade do arquivo, melhor ranqueado ele será em referência aos demais similares as preferências do usuário. Por exemplo, caso o sistemas de recomendação encontre dois arquivos "A" e "B", estes similares às preferências do usuário e com *score* de similaridade igual a "arquivo A = 0.8" e "arquivo B = 0.5". Neste cenário, o "arquivo A" será melhor ranqueado que o "arquivo B" no critério similaridade. A similaridade, torna-se imprescindível neste modelo de recomendação. Para tanto, este critério objetiva atender as preferências dos usuários em relação à filtragem de conteúdo relevante.

Critério Disponibilidade: refere-se ao tempo em que um arquivo estará disponível para o usuário. A disponibilidade, neste modelo, é medida em horas, ou seja a quantidade de horas em que um arquivo a ser recomendado está disponível na nuvem. Um arquivo só deve ser recomendado ao usuário se o mesmo estiver disponível e possibilitando o seu *download*. O critério disponibilidade representa uma das principais características e anseios quanto à tecnologia de computação em nuvem. A maioria dos usuários que migram para a nuvem são atraídos pela oferta de altas taxas de disponibilidade, elasticidade e mobilidade, que torna possível armazenar arquivos em grande quantidade e que estejam disponíveis e acessíveis a qualquer momento a partir da conexão com a Internet [7] [10] [23]. O cálculo do critério disponibilidade é apresentado na Equação 2:

$$D = h \cdot \left(\frac{1}{n}\right) \quad (2)$$

Na Equação 2, "D" é a quantidade horas em que um arquivo está disponível na nuvem, e "n" representa a quantidade de horas em que um arquivo pode ficar disponível na nuvem. Caso a nuvem fique disponivel durante todo o dia, "n" será igual a 24 (vinte e quatro) horas. A quantidade de horas de disponibilidade é normalizada em um valor entre 0 (zero) e 1 (um). O exemplo a seguir, demonstra como o critério disponibilidade contribui para a geração de uma recomendação. Considere que dois arquivos "A" e "B" são similares, o arquivo "A" está disponível na nuvem no intervalo de tempo (14 às 16 horas), totalizando duas horas de disponibilidade. O arquivo "B" está disponível na nuvem de (14 às 18 horas), totalizando quatro horas de disponibilidade. Desta maneira, o arquivo que será melhor ranqueado é o arquivo "B", por estar disponível na nuvem por um tempo superior que o arquivo "A", permitindo o seu *download* em um espaço de tempo maior. O objetivo central é mínimizar o risco do usuário não poder realizar o *download* e garantir que um arquivo recomendado esteja sempre acessível ao usuário.

Critério Taxa de *Download*: refere-se à taxa disponível para a realização do *download* de um arquivo na nuvem. O objetivo, é que arquivos que proporcionam melhores condições para a redução no tempo gasto no *download* sejam melhor ranqueados que os demais arquivos. A contribuição deste critério na redução do tempo gasto no *download* de um

arquivo recomendado, é produzida em conjunto com o critério "Tamanho do Arquivo", apresentado no próximo item. Por exemplo, no caso de termos dois arquivos similares às preferências do usuário, onde o arquivo "A" tem o seu tamanho igual a 10 (dez) *Gigabytes*, e o arquivo "B" tem o seu tamanho igual a 2 (dois) *Gigabytes*. Neste cenário, o arquivo "A" será melhor ranqueado que o arquivo "B", por proporcionar uma maior economia no tempo gasto em seu *download*. A taxa de *download* pode modificar o ranque de recomendações dependendo do momento em que a recomendação for calculada, principalmente em ambientes onde a taxa de *download* é oscilante. Este critério, tem valor de 0 (zero) a 3 (três) *Megabits* por segundo, este valor representa a media global de taxa de *downloads* apresentada pela Akamai [3]. Este critério é calculado pela Equação 3:

$$Td = ns \cdot \left(\frac{1}{n}\right) \tag{3}$$

Na Equação 3, a "Taxa de *Download*" é representada por "Td", onde "ns" representa à taxa de *downlaod* em Mbps, em seguida este valor é normalizado em um valor entre 0 (zero) e 1 (um), onde "n" representa o valor da media global de taxa de *downloads* em Mbps.

Critério Tamanho do Arquivo: este critério, corresponde ao tamanho do arquivo candidato a ser recomendado, tem como objetivo contribuir na tarefa de amenizar o tempo gasto no *downlaod* de um arquivo recomendado. Como explicado, no item critério "Taxa de *Download*", o critério "Tamanho do Arquivo", está diretamente relacionado com o critério que mensura a taxa de *download* disponível. O ranque de recomendação, mudará de acordo com a taxa disponível para *download*. No caso, da taxa de *download* ser baixa, os arquivos com tamanho menores devem ser mais bem ranqueados que seus similares que são maiores. Da mesma forma, quando a taxa de *download* é alta, os arquivos com tamanhos maiores devem ser mais bem ranqueados. Exemplificando, o ranqueamento deste critério considerando que um arquivo "A" é similar ao arquivo "B", o arquivo "A" têm o tamanho igual a 9 (nove) *gigabytes*. O arquivo "B", tem tamanho igual a 2 (dois) "*gigabytes*". Desta forma, o arquivo "B" será melhor ranqueado por apresentar melhores condições para a realização do seu *download* (menor tamanho), considerando que a taxa de *download* seja baixa. O cálculo deste critério é realizado pela Equação 4:

$$S = T \cdot \left(\frac{1}{n}\right) \tag{4}$$

Na Equação 4, o critério "Tamanho do Arquivo" é representado por "S". O tamanho do arquivo, é medido em *gigabytes*, pelo fato de que boa parte dos sistemas de armazenamento em nuvem limita o tamanho máximo de um arquivo que pode ser salvo em nuvem e o espaço disponível para o usuário no sistema em *gigabytes*. O tamanho do arquivo é multiplicado por $\frac{1}{n}$, para que seja normalizado por um valor de 0 (zero) a 1 (um), o valor 1 (um) é divido por n que é o tamanho máximo de um arquivo aceito no sistema de armazenamento em nuvem utilizado para implantação do modelo.

Critério Popularidade: este critério, representa a importância social de um arquivo na nuvem, avaliado por meio da quantidade de *downloads* que foram realizados de um mesmo arquivo. Quanto maior é a quantidade de *downloads* realizados de um arquivo, maior será a popularidade desse arquivo na rede, resultando em um melhor ranqueamento do mesmo. A seguir um exemplo sobre o ranqueamento deste critério: um arquivo "A" é similar ao arquivo "B", o arquivo "A" já teve dez *downloads* realizados, e o arquivo "B" já teve dezesseis *downloads* realizados. Desta forma, o aquivo "B" será melhor ranqueado, por obter um número maior de *downloads* efetuados no sistema de armazenamento em nuvem que o arquivo "A". O cálculo, deste critério é representado pela Equação 5:

$$R = Qd \cdot \left(\frac{1}{n}\right) \tag{5}$$

Na Equação 5, o critério "Popularidade do Arquivo" é representado por R. A cada *download* realizado de um determinado arquivo, o contador de *downloads* desse arquivo é incrementado em 1 (um). Este valor, é medido de 0 (zero) a n, onde n é a maior quantidade de *downloads* realizados em um único arquivo no sistema. O valor de n é obtido na observação do histórico de *downloads* de arquivos no sistema de armazenamento em nuvem. No cálculo do critério, a quantidade de *downloads* de um arquivo Qd é normalizada, multiplicando Qd por $\frac{1}{n}$, desta forma o valor resultante deste critério será entre 0 (zero) e 1 (um).

Tabela 1: Pesos dos Critérios

Critério	Peso
Similaridade	4
Disponibilidade	2
Taxa de *Download*	2
Tamanho do Arquivo	1
Popularidade	1

4.1 Pesos dos Critérios

Em um mecanismo de recomendação, os critérios devem ser ponderados por pesos, para compor o *score* de recomendação, resultando em um *ranking* com os itens que devem ser recomendados ao usuário. A partir, da realização de testes de execução com diferentes pesos no cálculo de recomendação que será apresentado na próxima seção, propomos a utilização dos pesos descritos na Tabela 1. Os testes realizados tiveram como objetivo verificar a variação no resultado do *score* final de recomendação e quais pesos apresentaria a menor variação no *score*. Desta forma, os arquivos recomendados não apresentariam uma grande variação de similaridade com as preferências dos usuários.

Critério Similaridade: tem peso maior peso que os demais critérios, seu objetivo é garantir que o conteúdo de um arquivo recomendado ao usuário seja similar as suas preferências. Outro ponto motivador para este critério corresponder a 40% do *score* de recomendação é amenizar ou solucionar um dos principais problemas da técnica de recomendação baseada em conteúdo: a sugestão de itens sempre muito parecidos, limitando os usuários de conhecer novos conteúdos [26] [30]. Desta forma, o modelo de recomendação atenderá as preferências do usuário e ao mesmo tempo estará recomendando novos conteúdos que são relacionados às preferência do usuário. O critério Similaridade, é medido de 0 (zero) a 1 (um), um arquivo que possua similaridade

[3]http://www.akamai.com/

igual a 0 (zero) em comparação as preferências do usuário, somente será recomendado caso ele tenha uma alta taxa de popularidade na rede, e mesmo assim, o arquivo não será bem ranqueado em relação aos demais que apresentem alguma similaridade com as preferências do usuário.

Critério Disponibilidade: tem peso 2 (dois), por ser um dos critério mais importante no modelo proposto, representa o tempo em que um arquivo está disponível na nuvem, tornando possível o *download* de um arquivo recomendado. Este critério, é essencial para a recomendação de arquivos baseada em características da nuvem, por representar uma das principais característica e vantagens da utilização de sistemas de armazenamento em nuvem. O valor do critério Disponibilidade, será de 0 (zero) a 1 (um), um arquivo somente poderá ser recomendado ao usuário, se o mesmo estiver disponível.

Critério Taxa de *Download*: Este critério, terá o seu valor medido de 0 (zero) a 1 (um). Um arquivo que possua uma baixa taxa de *download* e seu tamanho seja maior que o tamanho dos demais arquivos similares a ele, o seu *score* de recomendação será menor, e consequentemente ele não será tão bem ranqueado quanto seus similares, por que o seu processo de *download* demandará mais tempo e processamento. Um arquivo com baixa taxa de download poderá aparecer no topo do *ranking* de recomendação, desde que seu tamanho seja proporcional à baixa taxa de *download*. Para que um arquivo na nuvem se torne recomendável, esse critério deve ser maior que 0 (zero), desta forma será possível realizar o *download* do arquivo.

Critério Tamanho do Arquivo: Este critério, tem peso inferior aos demais critérios, por não ser um critério crítico. Assim, um arquivo que tenha o tamanho igual ao máximo aceito pelo ambiente, poderá ser recomendado se à taxa de *download* for alta, garantindo bom desempenho no *download* do arquivo.

Critério Popularidade: Este critério, tem o seu peso inferior aos demais critérios, por não ser um critério crítico. Portanto, um arquivo que não seja popular na nuvem poderá ser recomendado ao usuário, o mesmo ocorre com os arquivos novos na rede, se o arquivo for bem ranqueado nos outros critérios do modelo.

5. AVALIAÇÃO

Para a avaliação do modelo proposto, foi implementado um protótipo como parte do sistema de armazenamento em nuvem Ustore[4]. Foram implementados, os critérios propostos no modelo RecCloud e o cálculo de recomendação. O Ustore, é uma ferramenta de armazenamento em nuvem baseada em uma arquitetura P2P híbrida que tem como objetivo armazenar dados com baixo custo e de forma que os mesmos não se tornem indisponíveis com eventuais problemas na rede [13] [29].

O mecanismo de recomendação do Ustore é baseado em conteúdo. No Ustore, as preferências dos usuários são representadas por arquivos que o usuário tenha adicionado em sua conta no sistema. O sistema de recomendação, apresenta recomendações baseadas na similaridade entre o arquivo de preferência do usuário e os arquivos armazenados na nuvem. Para que o usuário receba uma recomendação, ela deve ser solicitada a partir de um arquivo em sua conta.

[4]http://usto.re/

5.1 Detalhes da Implementação

Critério Similaridade: a similaridade é calculada a partir do conteúdo dos arquivos, que são extraídos pelo Apache Lucene versão 3.6[5] e o Apache Tika versão 1.2[6].

Critério Disponibilidade: no Ustore, cada cliente possui um horário de funcionamento determinado inicialmente, que é utilizado para garantir a disponibilidade. Para chegar ao valor da taxa de disponibilidade de um cliente, subtraímos o tempo total possível para um cliente estar disponível em um dia (24 horas), pelo tempo em que o cliente ficou *on-line*. Desta forma, obtemos a quantidade de horas em que um cliente esteve disponível durante o dia

Critério Taxa de *Download*: é obtida a partir da observação do *download* de um arquivo qualquer, a partir de informações sobre o tempo gasto no *download* do arquivo e seu tamanho, chegamos à taxa de *download* da rede em "KBps" *Kilobits por segundo*. Para termos de avaliação, a taxa máxima de *download* utilizada foi a média global de taxa de *downloads* 3 (três) *Megabits* por segundo. Caso à taxa do usuário apresente um valor maior que a média global, o critério "Taxa de *Download*" será igual a 1 (um), representa que este critério recebeu o valor máximo no cálculo da recomendação.

Critério Tamanho do Arquivo: é obtido na base do Ustore em *KiloBytes* e convertido em *GigaBytes*. Para termos de avaliação da proposta, o tamanho máximo de um único arquivo no ambiente foi estabelecido em 10 *GigaBytes*. Desta forma, se um arquivo apresentar tamanho superior ao limite máximo estabelecido, este receberá o valor 1 (um), que representa a maior taxa ponderada do critério "Tamanho do Arquivo".

Critério Popularidade do Arquivo: é representada pela quantidade de vezes em que foram realizados *download* de um arquivo. No Ustore, a quantidade de vezes em que foi realizado o *download* de um arquivo pode ser obtida diretamente na base de dados do sistema. Para termos de avaliação desta proposta, atribuímos a quantidade de 10 *downloads*, como a maior quantidade de *downloads* realizados de um único arquivo na rede.

5.2 Detalhes da Avaliação

Para avaliação deste trabalho, foi utilizada uma base de dados composta por 1.400 (mil e quatrocentos) artigos acadêmicos. Para a realização dos testes, foi necessário que os arquivos recebessem uma classificação que indisse quais são os mais relevantes para o ranque de recomendação que será gerado. Para fazer essa classificação no conjunto de dados, foi utilizada a similaridade entre o contexto dos artigos com o conteúdo dos artigos utilizados para representar as preferências do usuário. Desta forma, todos os artigos que forem relacionados ao termo sistemas de recomendação, foram considerados relevantes para o ranque de recomendação. O resultado, foi de 462 (Quatrocentos e sessenta e dois) artigos foram considerados relevantes para esta avaliação, representando 33% da base de dados utilizada na realização dos testes.

Na avaliação, foi utilizada a metodologia proposta por Jain [18], onde é defendido que para realizar uma avaliação é preciso definir objetivos, métricas, fatores e níveis [22]. As métricas escolhidas para avaliar o desempenho do modelo

[5]https://lucene.apache.org/core/
[6]http://tika.apache.org/

RecCloud, estão descritas a seguir:

Tempo gasto no *Download*: foram efetuados *downloads* dos arquivos recomendados e medido o tempo gasto para realizar cada *download*. Os resultados obtidos, serão comparados com o tempo gasto no *download* de arquivos recomendados utilizando o modelo de recomendação baseado em conteúdo do Ustore.

Precisão: é a taxa de itens relevantes recomendados no resultado. É dada, através da proporção entre o número de arquivos relevantes recomendados e o número total de arquivos recomendados [4] [21] [31]. O cálculo da métrica Precisão é apresentado na Equação 6

$$Precisao = \frac{\mid \{arqrelevantes\} \bigcap \{arqrecomendados\} \mid}{\{arqrecomendados\}} \quad (6)$$

Na Equação 6, "*arqrelevantes*" é a quantidade de arquivos recomendados que fazem parte do ranque de relevância, e "*arqrecomendados*" é a quantidade de arquivos recomendados para cada solicitação de recomendação. O resultado, é representado por valores entre 0 (zero) e 1 (um), quanto mais próximo de 1 (um), mais preciso é o sistema.

***Recall*:** é a taxa de itens relevantes recomendados em relação à quantidade total de itens relevantes [21]. O cálculo da métrica *Recall* é apresentado na Equação 7.

$$Recall = \frac{\mid \{arqrelevantes\} \bigcap \{arqrecomendados\} \mid}{\{arqrelevantes\}} \quad (7)$$

Na Equação 7, "*arqrelevantes*" é a quantidade de arquivos recomendados que fazem parte do ranque de relevância, e "*arqrecomendados*" é a quantidade de arquivos recomendados para cada solicitação. O resultado é representado por valores entre 0 (zero) e 1 (um), quanto mais próximo de 1 (um) mais o sistema satisfaz a solicitação da recomendação.

***F-measure*:** é a média ponderada das taxas de Precisão e *Recall*. O cáculo da métrica *F-measure* é apresentado na Equanção 8

$$F - measure\ \alpha = \frac{(1+\alpha) \cdot precisao \cdot recall}{(\alpha \cdot precisao) + recall} \quad (8)$$

Nesta avaliação, as taxas de Precisão e *Recall* têm o mesmo fator de importância. desta forma, o valor de α é igual a 1 (um). Logo, esta função só retornará um valor no intervalo entre 0 (zero) e 1 (um). Como parâmetros para comparação, foram consideradas satisfatórias taxas de Precisão de 0.40 e *Recall* de 0.42, resultados semelhantes obtidos por Blank et al. [5], Tanaka et al. [31] e Zhang et al. [33].

5.3 Cénarios de Availiação

A avaliação deste trabalho, foi realizada em um sistema real de armazenamento em nuvem (Ustore). Na execução, foram montados dois cenários da seguinte forma:

Cenário I: O objetivo foi avaliar o desempenho do modelo RecCloud. Foram solicitadas recomendações para dez diferentes arquivos e avaliada a quantidade de arquivos recomendado para cada solicitação. Foram avaliadas as métricas Precisão, *Recall* e *F-Measure* observando os arquivos recebidos como recomendação. Os ranques de recomendação analisados, foram divididos em três níveis, o primeiro nível retornando 5 (cinco) arquivos, o segundo nível retornando 10 (dez) arquivos e nível três retornando 15 (quinze) arquivos como recomendação. Foram solicitadas recomendações para 10 (dez) artigos no modelo RecCloud, estas solicitações resultaram respectivamente em 50 (cinquenta), 100 (cem) e 150 (cento e cinquenta) artigos recomendados.

Tabela 2: Resultados do cenário I

	Nível I	Nível II	Nível III
Precisão	0.68	0.46	0.44
Recall	0.21	0.29	0.42
F-measure	0.32	0.35	0.42

Cenário II: o objetivo foi medir o tempo gasto no *download* de arquivos recomendados e comparar os resultados com um modelo baseado em conteúdo. Com isso, foi possível avaliar se o modelo proposto atingiu um dos objetivos, que é amenizar o tempo gasto no *download* dos arquivos recomendados. Neste cenário, avaliamos a métrica "Tempo Gasto no *Download*", foram solicitadas recomendações para 10 (dez) artigos diferentes no modelo RecCloud e no modelo baseado em conteúdo do Ustore, estas solicitações resultaram em 100 (cem) artigos recomendados, 50 (cinquenta) por cada modelo. Nesta métrica, foram utilizados ranques de recomendação com 5 (cinco) artigos retornados para cada solicitação de recomendação.

5.4 Resultados

Cenário I: Na Tabela 2, são apresentados os resultados obtidos nas taxas de Precisão, *Recall* e *F-measure*, estes resultados foram alcançados utilizando o modelo RecCloud para cada solicitação de recomendação e variando a quantidade de arquivos recomendados para cada solicitação nos nívelis I, II e III.

A partir dos resultados apresentados na Tabela 2, foi observado que a melhor taxa de precisão obtida foi de 0.68 no "nível I", onde foram retornados 5 (cinco) artigos para cada solicitação de recomendação. Os resultados obtidos na métrica de precisão foram superiores aos resultados utilizados como referência para termos de comparação. A melhor taxa de *recall* obtida foi de 0.42 no "nível III", onde foram retornados 15 (quinze) artigos para cada solicitação de recomendação. A melhor taxa de *Recall* obtida nesta avaliação foi similar à taxa obtida no trabalho de Zhang et al. [33]. A partir das taxas de Precisão e *Recall*, foi calculada a taxa de *F-Measure*, onde foi obtido como melhor taxa o valor 0.42, este valor foi obtido no "nível III".

Analisando os resultados apresentados, foi observado que o melhor resultado da taxa de Precisão foi obtido em nível diferente do nível onde foi obtida a melhor taxa de *Recall*, e que, a maioria dos arquivos relevantes recomendados estava no início dos ranques de recomendação. Este cenário, justifica-se pelo critério de similaridade, que representa 40% de cada recomendação. Desta forma, os artigos com maiores taxas de similaridade ficaram no início dos ranques e os artigos que mesmo tendo a sua taxa de similaridade baixa foram classificados como recomendáveis e foram recomendados no final dos ranques de recomendação.

A partir da comparação dos resultados apresentados no decorrer deste trabalho com resultados encontrados na literatura, como, por exemplo, os trabalhos de Blank et al. [5], Tanaka et al. [31] e Zhang et al. [33], pode-se concluir que a proposta apresentada por este trabalho obteve resultados satisfatórios na geração de recomendações em sistemas de armazenamento em nuvem, utilizando características do

Tabela 3: Resultados do cenário II

	Média	Máximo	Mínimo
RecCloud	900	2.200	400
RecUstore	1.150	2.700	150

ambiente associadas à técnica de recomendação baseada em conteúdo.

Cenário II: neste cenário, foram realizadas dez solicitações de recomendações para 10 (dez) artigos diferentes no modelo RecCloud e no modelo baseado em conteúdo do Ustore, estas solicitações resultaram em 100 (cem) artigos recomendados 50 (cinqueta) arquivos recomendados por cada modelo. Foi analisado, o tempo gasto no *download* de cada arquivo recomendado. Na Tabela 3, é apresentada uma comparação do resultado obtidado em cada modelo (média, máximo e o mínimo de tempo gasto nos *downloads* realizados).

A partir dos resultados apresentados na Tabela 3 em milissegundos, foi observado que, o modelo proposto neste trabalho proporcionou redução no tempo gasto no *download* dor arquivos recomendados. A redução média de tempo gasto nos *downloads* foi de 207,06 milissegundos, o que representa uma redução de 17,8%.

6. CONCLUSÃO

A avaliação realizada neste trabalho obteve resultados próximos aos resultados encontrados na literaturado sobre avaliação de sistemas de recomendação. Desta forma, é possível consideração como relevantes às contribuições apresentadas por esta pesquisa. O escopo deste trabalho, não abrange todas as possíveis características da nuvem e de sistemas de armazenamento em nuvem que podem ser utilizadas na geração de recomendações. Desta forma, como trabalhos futuros podem ser adicionados novos critérios ao modelo de recomendação, assim como, realizar a avaliação do modelo com outras técnicas de avaliação de sistemas de recomendação.

7. REFERÊNCIAS

[1] L. M. d. L. A. V. F. P. R. T. e. A. C. S. Adriano de Oliveira Tito, Arley Ramalho Rodrigues Ristar. Recroute: Uma proposta de aplicativo para recomenda de rotas de nibus utilizando informas contextuais dos usuos. *Anais do X Simp Brasileiro de Sistemas de Informas*, pages 218–223, 2014.

[2] T. B. M. d. S. Alezy Oliveira Lima, Ricardo Alexandre Afonso. Plataforma pguide: um modelo de recomenda para usuos ms. *Anais do X Simp Brasileiro de Sistemas de Informa*, pages 73 – 84, 2013.

[3] A. Ansari, S. Essegaier, and R. Kohli. Internet recommendation systems. *JOURNAL OF MARKETING RESEARCH*, 37(3):363–375, 2000.

[4] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[5] I. Blank, L. Rokach, and G. Shani. Leveraging the citation graph to recommend keywords. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 359–362, New York, NY, USA, 2013. ACM.

[6] R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[7] J. F. S. Carvalho. Um mapeamento sistematico de estudos em cloud computing. Master's thesis, Universidade Federal de Pernambuco (UFPE), 2012.

[8] Y.-C. Chen, H.-C. Huang, and Y.-M. Huang. Community-based program recommendation for the next generation electronic program guide. *Consumer Electronics, IEEE Transactions on*, 55(2):707–712, 2009.

[9] C. M. R. da Silva, J. L. C. da Silva, R. M. Melo, R. B. Rodrigues, L. R. Lucien, S. P. D. Melo, A. Colares, and V. C. Garcia. A privacy maturity model for cloud storage services. In *2014 IEEE 7th International Conference on Cloud Computing, Anchorage, AK, USA, June 27 - July 2, 2014*, pages 944–945, 2014.

[10] C. M. R. da Silva, J. L. C. da Silva, R. B. Rodrigues, G. M. M. Campos, L. M. Nascimento, and V. C. Garcia. Security threats in cloud computing models: Domains and proposals. In *2013 IEEE Sixth International Conference on Cloud Computing, Santa Clara, CA, USA, June 28 - July 3, 2013*, pages 383–389, 2013.

[11] C. M. R. da Silva, J. L. C. da Silva, R. B. Rodrigues, L. M. Nascimento, and V. C. Garcia. Systematic mapping study on security threats in cloud computing. *CoRR*, abs/1303.6782, 2013.

[12] J. Deng, J. Hu, A. Liu, and J. Wu. Research and application of cloud storage. In *Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on*, pages 1–5, 2010.

[13] R. F. A. F. J. G. V. T. F. Dur F. Assad. Usto.re: A private cloud storage software system. In F. Daniel, P. Dolog, and Q. Li, editors, *Web Engineering*, volume 7977 of *Lecture Notes in Computer Science*, pages 452–466. Springer Berlin Heidelberg, 2013.

[14] N. C. L. R. e Edmir Parada Vasques Prado. Caractericas dos servi de computa em nuvem usados por organizas brasileiras. *Anais do X Simp Brasileiro de Sistemas de Informa*, pages 661 – 673, 2013.

[15] P. L. d. G. e. J. P. d. A. Fernando M. Figueira Filho. Um sistema de recomenda para fs de discussa web baseado na estimativa da expertise e na classifica colaborativa de conteudo. *Anais do V Simp Brasileiro de Sistemas de Informa*, pages 306 – 313, 2007.

[16] J. Gantz and D. Reinsel. Extracting value from chaos state of the universe : An executivesummary., Junho 2011. 1-12.

[17] J. L. Herlocker. *Understanding and improving automated collaborative filtering systems*. PhD thesis, University of Minnesota, 2000. AAI9983577.

[18] R. K. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley, 1 edition, Apr. 1991.

[19] G. Jung, T. Mukherjee, S. Kunde, H. Kim, N. Sharma, and F. Goetz. Cloudadvisor: A recommendation-as-a-service platform for cloud configuration and pricing. In *Services (SERVICES), 203 IEEE Ninth World Congress on*, pages 456–463, 2013.

[20] C.-F. Lai, J.-H. Chang, C.-C. Hu, Y.-M. Huang, and

H.-C. Chao. Cprs: A cloud-based program recommendation system for digital tv platforms. *Future Gener. Comput. Syst.*, 27(6):823–835, June 2011.

[21] S. Lee, D. Lee, and S. Lee. Personalized dtv program recommendation system under a cloud computing environment. *IEEE Trans. on Consum. Electron.*, 56(2):1034–1042, May 2010.

[22] M. A. S. Machado. Uma abordagem para indexacao e buscas full-text baseadas em contedo em sistemas de armazenamento em nuvem. Master's thesis, Universidade Federal de Pernambuco (UFPE), 2013.

[23] P. Mell and T. Grance. The NIST Definition of Cloud Computing. Technical report, National Institute of Standards and Technology, Information Technology Laboratory, July 2009.

[24] M. J. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27(3):313–331, 1997.

[25] W. O. F. G. M. M. C. V. C. G. F. A. D. R. E. A. Ricardo Batista Rodrigues, Carlo M. R. da Silva. A cloud-based recommendation system. In B. White and P. Isaías, editors, *Proceedings of the IADIS International Conference WWW/Internet 2013*, pages 384 – 386, Fort Worth, Texas, USA, October 2013. 978-989-8533-16-6.

[26] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[27] R. B. Rodrigues. Reccloud: Um modelo de recomenda de arquivos para sistemas de armazenamento em nuvem. Master's thesis, Universidade Federal de Pernambuco (UFPE), 2014.

[28] R. B. Rodrigues, F. A. Dur V. C. Garcia, C. M. R. da Silva, R. R. Souza, and R. E. Assad. A cloud-based recommendation model. In *7th Euro American Conference on Telematics and Information Systems, EATIS'14, Valparaiso, Chile - April 02 - 04, 2014*, page 23, 2014.

[29] M. S. P. J. T. G. V. A. R. Silva, A. Machado. Desenvolvendo aplicativos peer-to-peer (p2p) no contexto de data storage para ambientes de cloud computing. *Computa, S. B., editor, SBSI.*, 2013.

[30] N. Stormer, H.; Werro and D. Risch. Recommending products with a fuzzy classification. *CollECTeR Europe*, 2006.

[31] M. G. M. Thiago Fujisaka Tanaka. Classifica de revises para constru de perfm sistemas de recomenda. In *Webmedia 2012 XVIII Simp Brasileiro de Sistemas Multima e Web*, 2012.

[32] W. Zeng, Y. Zhao, K. Ou, and W. Song. Research on cloud storage architecture and key technologies. In *Proceedings of the 2Nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, ICIS '09, pages 1044–1048, New York, NY, USA, 2009. ACM.

[33] Z. Zhang, S. Shang, S. R. Kulkarni, and P. Hui. Improving augmented reality using recommender systems. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 173–176, New York, NY, USA, 2013. ACM.