Ferramentas para a detecção de grupos em WIKIS

Fabrício J. Barth
Centro Universitário Senac
Av. Eng. Eusébio Stevaux, 823
Santo Amaro, São Paulo - SP
fabricio.jbarth@sp.senac.br

***Resumo*—Este artigo explora duas técnicas, agrupamento hierárquico de documentos e análise de grafos, para a detecção de grupos em WIKIS. Ambas as abordagens exploradas neste trabalho fazem uso das informações sobre o histórico de criação e alteração de páginas em um WIKI. Para a validação da proposta foi implementado um software[1] que: acessa o conteúdo produzido em WIKIS; gera os grafos e os agrupamentos hierárquicos de acordo com a solicitação dos usuários, e; mostra os resultados encontrados em uma interface gráfica. Os resultados encontrados mostram que este tipo de análise pode ser utilizada para a identificação de pessoas e grupos com interesses e habilidades similares.**

***Keywords*-Wiki, Algoritmos de Agrupamento, Análise de Redes Sociais**

I. INTRODUÇÃO

O termo WIKI é utilizado para identificar um tipo específico de coleção de documentos em hipertexto ou o software colaborativo usado para criá-lo. Um WIKI permite que os documentos sejam editados colaborativamente através da utilização de um navegador Web, permitindo a geração e distribuição rápida de conteúdo e a fácil colaboração entre usuários. Tipicamente, um WIKI permite que usuários possam: (i) adicionar novo conteúdo; (ii) ligar um conteúdo ao outro através de hiperlinks; (iii) editar conteúdo já existente; (iv) organizar e estruturar o conteúdo; (v) visualizar o conteúdo, e; (vi) acessar o histórico de contribuições. Adicionalmente, WIKIS usam vários mecanismos para monitorar o histórico de contribuições dos usuários.

As vantagens do uso de WIKI são: (i) WIKIS geram uma rede de conhecimento ligando pessoas ao conteúdo criado; (ii) WIKIS auxiliam na criação de documentos consensuais, e; (iii) através de WIKIS é fácil gerar informações a partir de diversas fontes. Nos últimos anos, inúmeras organizações vêm explorando WIKIS como ferramentas para: Gestão de Projetos, Gestão de Conhecimento, Levantamento de Requisitos junto ao cliente, Levantamento de Idéias entre os colaboradores, entre outras atividades. Entre as organizações que utilizam este tipo de ferramental estão: IBM, NASA, Google e Microsoft [4].

[1]Endereço para o vídeo de demonstração do protótipo que implementa os algoritmos analisados neste trabalho: http://www.youtube.com/watch?v=HaTTwErV3_g

Muitos autores têm como premissa que usuários de qualquer serviço sobre a Internet, incluindo WIKIS, constituem um grande grupo. Este grande grupo possui perfis similares e pode ser dividido em sub-grupos [5]. A descoberta destes sub-grupos pode facilitar a interação de pessoas com pessoas e de pessoas com sistemas computacionais de várias maneiras: através da recomendação de pessoas para a execução de atividades; para a criação de comunidades de prática; recuperação de informação otimizada para um determinado grupo de trabalho; personalização de serviços, e; identificação de pessoas com o mesmo interesse [1].

Uma pessoa que atua em um WIKI realiza diversas atividades simultaneamente, coopera com diversos colegas na execução de um objetivo comum, manipulando diversos documentos (fonte e produto) de informação ao longo do trabalho. Quando uma pessoa cria, remove ou altera qualquer tipo de conteúdo significa que esta pessoa, de alguma maneira, está conectada ao tema que aquele conteúdo faz parte.

Acredita-se que através da análise do conteúdo manipulado por diversas pessoas em um WIKI seja possível identificar grupos de pessoas com interesses, habilidades e interesses comuns. A identificação de grupos de pessoas com interesses, habilidades e interesses similares pode ser realizada de duas formas: (*i*) formando grupos a partir de pessoas que manipulam documentos considerados similares, e; (*ii*) formando grupos a partir de pessoas que contribuem para a construção dos mesmos documentos.

O objetivo deste trabalho é explorar duas técnicas, agrupamento hierárquico de documentos e análise de grafos, para a detecção de grupos em WIKIS. Com o agrupamento hierárquico de documentos pretende-se identificar grupos através de documentos similares e com a análise de grafos pretende-se identificar grupos que estão conectados através dos mesmos documentos.

II. DETECÇÃO DE COMUNIDADES EM GRAFOS

Um WIKI consegue manter o registro de alterações das páginas utilizando um sistema de controle de versões. Na maioria dos casos, um sistema de controle de versões mantem informações sobre quem alterou o texto, quando alterou e o que foi alterado. Estas informações são úteis

978-0-7695-4240-9/10 $26.00 © 2010 IEEE
DOI 10.1109/SBSC-II.2010.11

para determinar o relacionamento entre pessoas e documentos, pois possuem informações sobre quem colaborou para a confecção de determinado documento. Como exemplo, pode-se considerar um histórico de alterações de páginas como apresentado na tabela I.

Após uma análise deste histórico é possível identificar o grupo de pessoas que manipulou determinado grupo de documentos. Na figura 1, item *(a)*, é possível visualizar um esquema que representa esta interação de pessoas com documentos. Nesta figura é possível visualizar um conjunto de documentos $\{d_1, \cdots, d_5\}$ e um conjunto de pessoas $\{u_1, \cdots, u_{11}\}$ - exatamente os mesmos representados na tabela I. O item *(a)* da figura 1 mostra, por exemplo, que o documento d_1 é manipulado pelas pessoas u_1, u_2, u_3 e u_4. Ao mesmo tempo, mostra que a pessoa u_4 também ajuda a elaborar o documento d_2.

Esta associação entre pessoas e documentos, se tratada, gera uma outra informação que é a relação entre pessoas através da manipulação dos documentos. Por exemplo, o fato das pessoas u_5, u_6 e u_7 manipularem o mesmo documento d_2 faz acreditar que este é um grupo com interesses em comum. Assim como, o fato da pessoa u_7 também manipular o documento d_5 faz acreditar que existe uma relação entre o grupo $\{u_5, u_6, u_7\}$ e o grupo $\{u_7, u_{10}, u_{11}\}$ por causa da pessoa u_7.

Estas associações são utilizadas na criação de um grafo não direcionado $G(N, E)$, onde cada nodo em N representa uma pessoa que editou no mínimo um documento no WIKI. Um vérticie é adicionado ao conjunto E se, no mínimo, o documento representado pelo vértice foi editado por duas pessoas. Os grupos são definidos encontrando-se todos os sub-grafos biconectados em um grafo não direcionado g. No item (a) da figura 2 é possível visualizar o grafo gerado a partir da análise do histórico apresentado na tabela I.

III. AGRUPAMENTO HIERÁRQUICO DE DOCUMENTOS

Para identificar documentos que são similares em termos de conteúdo pode-se utilizar diversas técnicas. Entre as técnicas mais difundidas estão os algoritmos de agrupamento. O objetivo dos algoritmos de agrupamento é colocar documentos similares em um mesmo grupo e documentos não similares em grupos diferentes. Os algoritmos de agrupamento são divididos em dois tipos: aqueles que geram agrupamentos planos e os que geram agrupamentos hierárquicos [3].

Um agrupamento hierárquico é representado por uma árvore [3]. Os nós folhas são os documentos. Cada nó intermediário representa o agrupamento que contém todos os documentos de seus descendentes.

Este trabalho faz uso de agrupamentos hierárquicos de documentos para identificar grupos de usuários que manipulam documentos sintaticamente similares. Todos os usuários que contribuiram para o desenvolvimento de um documento são inseridos no grupo que o documento faz parte. Por exemplo, no item (b) da figura 2 é apresentado o agrupamento hierárquico para o exemplo apresentado no item *b* da figura 1. Neste agrupamento é possível visualizar uma relação entre os usuários u_{10}, u_{11}, u_8 e u_9 que manipulam documentos similares (d_5 e d_4).

Os usuários u_{10} e u_{11} manipulam o documento d_5 e os usuários u_8 e u_9 manipulam o documento d_4. Aparentemente, não existe ligação alguma entre os usuários u_{10}, u_{11}, u_8 e u_9. No entanto, a similaridade entre os documentos d_5 e d_4 pode indicar que existam interesses ou atividades em comum entre estes usuários.

Uma distinção entre a abordagem hierárquica e as demais é que o resultado obtido não é constituído apenas de uma partição do conjunto de dados inicial, mas sim de uma hierarquia que descreve um particionamento diferente a cada nível analisado. Um conjunto de dados contém, geralmente, diversos agrupamentos, que por sua vez, contém sub-agrupamentos. Os sub-agrupamentos podem ainda ser formados a partir do agrupamento de outros agrupamentos menores, e assim sucessivamente [3], [2].

Neste trabalho, isto também acontece com os agrupamentos dos usuários, pois os agrupamentos de usuários estão associados aos agrupamentos de documentos (ver item (b) da figura 2). Isto faz com que exista uma flexibilidade em relação à análise dos diferentes níveis de granularidade e densidade de agrupamentos. O principal uso deste algoritmo, neste domínio de aplicação, é realizar a análise exploratória dos documentos e usuários ao mesmo tempo.

IV. IMPLEMENTAÇÃO

Com o intuito de validar a proposta aqui apresentada, foi implementada uma ferramenta[2] que acessa o conteúdo produzido em WIKIS através de uma interface XML-RPC[3]. Esta ferramenta é capaz de conectar-se e acessar o conteúdo de diversos WIKIS, desde que a interface XML-RPC implementada pelo WIKI respeite o padrão do *wiki Confluence*[4]. Os principais testes foram realizados com o projeto *Xwiki*[5].

Foram acessados ambientes colaborativos de escrita utilizados por aproximadamente 300 pessoas em aproximadamente 15 projetos distintos. Para a detecção e apresentação dos grupos, foi desenvolvido uma aplicação desktop utilizando a linguagem de programação Java.

A. Identificação de grupos em grafos

Para a construção do grafo foi implementada uma rotina que identifica todas as pessoas que contribuiram para cada um dos documentos do WIKI. Estas pessoas são inseridas em uma lista de nodos N. Cada nodo N em $G(N, E)$ representa uma única pessoa e cada pessoa está representada por um

[2]Endereço para o vídeo de demonstração do protótipo implementado: http://www.youtube.com/watch?v=HaTTwErV3_g

[3]http://en.wikipedia.org/wiki/Xml-rpc

[4]http://www.atlassian.com/software/confluence/

[5]http://www.xwiki.org

Tabela I
EXEMPLO DE HISTÓRICO DE CRIAÇÃO E ALTERAÇÃO DE PÁGINAS EM UM WIKI

Documento	Versão	Editor	Data	Documento	Versão	Editor	Data
d_1	1	u_1	$\cdots$	d_2	4	u_7	$\cdots$
d_1	2	u_2	$\cdots$	d_3	1	u_5	$\cdots$
d_1	3	u_2	$\cdots$	d_3	2	u_6	$\cdots$
d_1	4	u_3	$\cdots$	d_3	3	u_6	$\cdots$
d_1	5	u_4	$\cdots$	d_4	1	u_8	$\cdots$
d_2	1	u_4	$\cdots$	d_4	2	u_9	$\cdots$
d_2	2	u_5	$\cdots$	d_5	1	u_{10}	$\cdots$
d_2	3	u_6	$\cdots$	d_5	2	u_{11}	$\cdots$

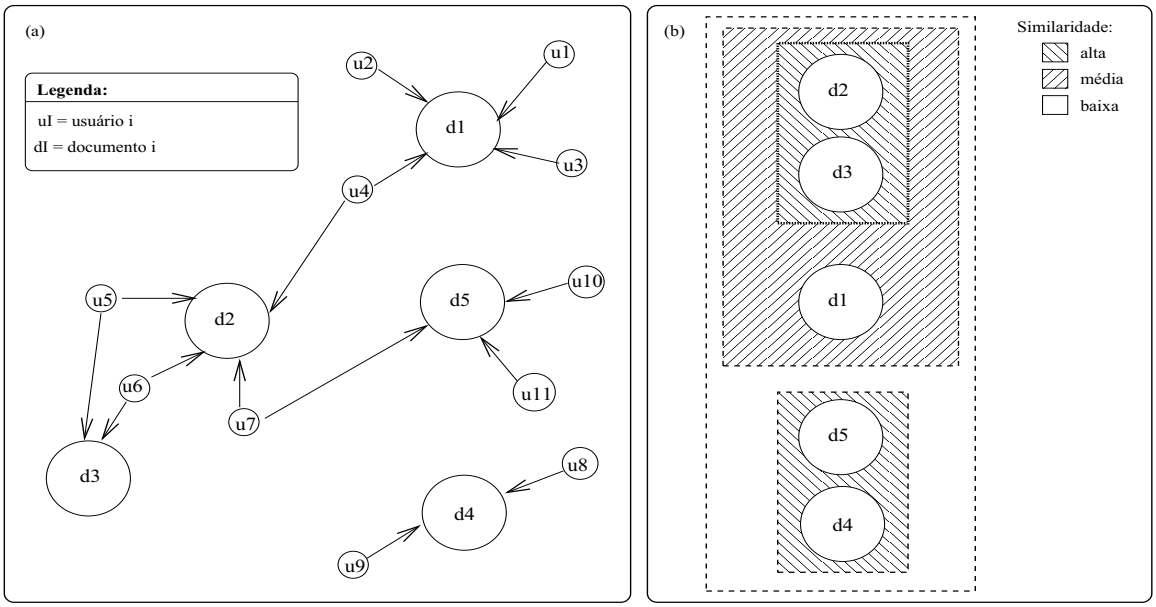

Figura 1. *(a)* Exemplo da ligação entre usuários e documentos. *(b)* Exemplo da relação de similaridade entre documentos

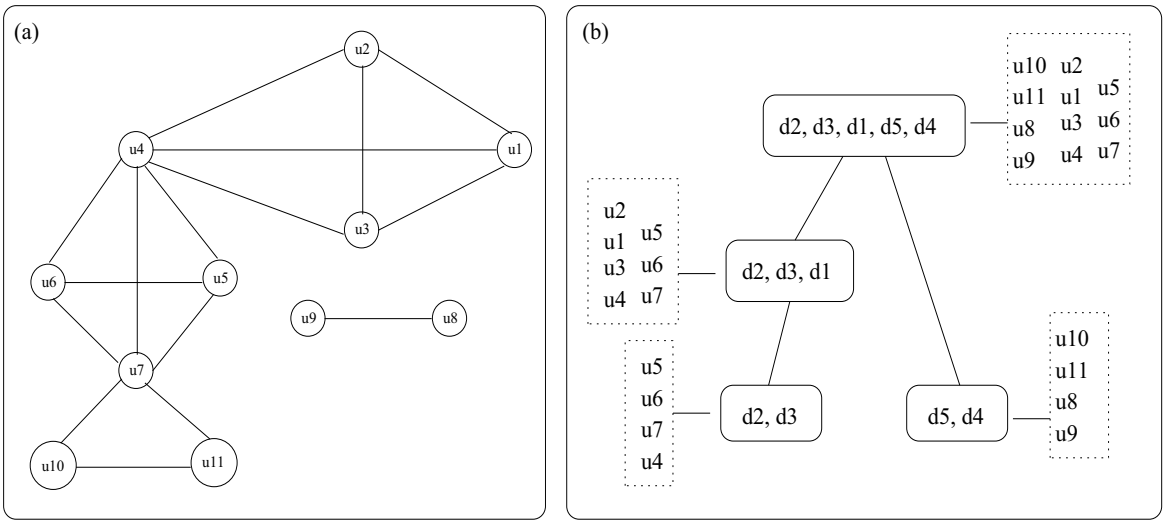

Figura 2. (a) Grafo gerado a partir da análise do histórico apresentado na tabela I. (b) Agrupamento hierárquico referente ao exemplo apresentado no item *b* da figura 1

único nodo N. Para cada pessoa em N são identificadas as arestas E que ligam esta pessoa a outras pessoas que manipularam os mesmos documentos.

Para a visualização dos grafos foi utilizada a API Jung[6]. Esta API já implementa diversas funcionalidades que facilitam a manipulação e visualização de grafos. O algoritmo utilizado para a identificação dos sub-grupos no grafo foi o algoritmo *VoltageClusterer* [8]. Este algoritmo permite descobrir sub-grupos em um grafo em tempo que cresce linearmente de acordo com o tamanho do grafo.

No protótipo implementado, o usuário pode escolher (clicar) em qualquer nodo e a partir desta ação obter informações sobre a pessoa selecionada, tais como: com quem ela está conectada e que documentos que está pessoa ajudou a construir.

B. Identificação de grupos em agrupamentos hierárquicos

Para a implementação do agrupamento hierárquico foi utilizado um algoritmo da classe *Unweighted Pair Group Method with Arithmetic mean - UPGMA* [2], [3]. Trata-se de um algoritmo *bottom-up* que usa uma função de distância euclidiana como função de similaridade. O algoritmo recebe um conjunto de documentos, iniciando com um agrupamento para cada documento. Em cada passo, os dois agrupamentos mais similares são determinados e unidos em um novo agrupamento. Elimina-se os dois agrupamentos mais similares e adiciona-se o novo agrupamento ao conjunto de agrupamentos. O algoritmo é finalizado quando o número de agrupamentos for igual a 1.

O pré-processamento de cada documento inclui algoritmos de *stemming* [6], lista de *stop-words* [3] e a transformação de cada documento em um vetor utilizando a equação *TF-IDF* [7]. O algoritmo utilizado para cálculo do agrupamento hierárquico é um algoritmo com ordem de grandeza $O(n^2)$, onde n é o número de uniões realizadas durante o processo do agrupamento hierárquico. O número de uniões é exatamente o número de documentos utilizados menos um ($docs - 1$). Em uma máquina com processador Pentium 3 e 512 MB de memória, cada etapa do algoritmo para agrupamento hierárquico é processada em aproximadamente 1 milisegundo. Nesta situação, o tempo total de processamento de um agrupamento com 200 documentos é de 40 segundos.

V. Considerações finais

O desenvolvimento deste trabalho partiu da hipótese que a partir da análise do conteúdo manipulado por diversas pessoas em um WIKI seja possível identificar grupos de pessoas com interesses, habilidades e atividades em comum. Para isso, este trabalho explorou duas técnicas: agrupamento hierárquico de documentos e análise de grafos.

Ambas as abordagens exploradas neste trabalho fazem uso das informações sobre o histórico de criação e alteração de páginas em um WIKI. Para a implementação do agrupamento hierárquico de documentos foi utilizado o algoritmo UPGMA e para a implementação do identificador de grupos em grafos foi utilizado o algoritmo *VoltageClusterer*.

Para a validação da proposta foi implementado um software que: acessa o conteúdo produzido em WIKIS através de uma interface XML-RPC; gera os grafos e os agrupamentos hierárquicos de acordo com a solicitação dos usuários, e; mostra os resultados encontrados em uma interface gráfica. Os testes foram realizados com WIKIS utilizados por aproximadamente 300 pessoas em 15 projetos distintos. Os grupos encontrados utilizando a análise de grafos foram grupos muito similares aos formalmente definidos no ambiente de trabalho. No entanto, além dos grupos, foi possível identificar alguns *hubs* (pessoas que conectam grupos) durante a análise. Alguns dos resultados encontrados com a técnica de agrupamento hierárquico de documentos foram inesperados, pois, com esta técnica foi possível identificar grupos de pessoas com interesses e atividades em comum que nunca trabalharam nos mesmos projetos.

Os resultados encontrados mostram que este tipo de análise pode ser utilizada para a identificação de pessoas e grupos com interesses e habilidades similares. Em grandes corporações, isto pode ser utilizado para a implantação de comunidades de prática, para auxiliar na reestruturação de alguma área, em ações de recursos humanos (RH), entre outras atividades.

[6] http://jung.sourceforge.net/

Referências

[1] Tom Gross and Wolfgang Prinz. Modelling shared contexts in cooperative environments: Concept, implementation, and evaluation. *Computer Supported Cooperative Work*, 13(3-4):283–303, August 2004.

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

[3] Cristopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 2003.

[4] D. E. O'Leary. Wikis: from each according to his knowledge. *IEEE Computer Society*, pages 34–41, 2008.

[5] Jon Orwant. Heterogeneous learning in the doppelganger user modeling system. *User Modeling and User-Adapted Interaction*, 4(2):107–130, 1995.

[6] M. Porter. An algorithm for suffix stripping program. *Program*, 14(3):130–137, 1980.

[7] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

[8] Fang Wu and Bernardo Huberman. Discovering communities in linear time: A physics approach. *European Physics Journal*, pages 331–338, 2004.