Identificação de perfis de usuários a partir dos *logs* de acesso de um servidor Web

Fabrício Jailson Barth & Edson Satoshi Gomi
*Laboratório de Engenharia de Conhecimento (*KNOMA*)*
Escola Politécnica da Universidade de São Paulo
Email: {fabricio.barth;edson.gomi}@poli.usp.br

1 Objetivos

O objetivo deste trabalho é verificar a aplicabilidade de uma técnica de aprendizado de máquina não-supervisionada na identificação de perfis de usuários, provendo informação útil para a customização de serviços da Internet. Em particular, o maior interesse deste trabalho é construir modelos de comunidade/perfis [1] que representam padrões de uso de serviços que podem ser associados à diferentes tipos de usuários que acessam sites de laboratórios de pesquisa.

2 Método

O método utilizado neste trabalho consiste de três etapas: (1) a partir das informações contidas nos arquivos de *log* é formado o conjunto de exemplos de treinamento para o algoritmo de aprendizado. Esta etapa consiste em extrair IP, data e URL de cada acesso realizado. Agrupar os acessos em seções. Onde, cada seção corresponde a um conjunto de acessos realizados a partir do mesmo IP e dentro de uma faixa de tempo (i.e., 30 ou 60 minutos). Cada exemplo corresponde a uma seção e descreve a interação que o usuário teve com o site; (2) Aplicar sobre o conjunto de treinamento criado o algoritmo COBWEB[1], e; (3) Analisar o cluster hierárquico com o intuito de definir as comunidades/perfis e os seus respectivos padrões de comportamento.

3 Resultados

Nos experimentos realizados foram utilizados dados de acesso correspondentes à três meses. Como resultado do processo de transformação dos arquivos de *logs* no conjunto de treinamento, obteu-se 1138 exemplos. Para este conjunto de treinamento foi encontrado uma hierarquia com 789 *clusters*. No segundo nível da hierarquia formaram-se dois *clusters*. No primeiro *cluster*, com 676 objetos, identificou-se um comportamento similar ao das pessoas que visitam o site e acessam todas as páginas do site tentando conhecer mais sobre o grupo de pesquisa. No outro *cluster*, o comportamento identificado foi o de acesso direto à páginas específicas do site (i.e., tutoriais).

[1] O COBWEB [2] é um algoritmo incremental que executa uma busca para obter um *cluster* hierárquico

4 Considerações Finais

Apesar dos padrões de comportamento encontrados nos experimentos não serem muito significativos, acredita-se que a utilização do método mencionado neste trabalho fornece informações mais ricas do comportamento do usuário do que apenas análise estatística de um serviço - como normalmente é feito. Além disso, acredita-se que é possível utilizar os padrões encontrados para a definição de estereótipos. A maior dificuldade encontrada para realização deste trabalho foi a análise dos resultados. Infelizmente o resultado retornado pelo algoritmo COBWEB não é tão claro como a literatura informa. O trabalho de identificação de padrões acaba se transformando em um trabalho que exige um grande esforço manual. Para os próximos trabalhos é sugerido a adição de uma etapa de interpretação da hierarquia de clusters. Esta interpretação pode ser feita de maneira automática utilizando algum algoritmo de aprendizado indutivo.

Referências

[1] J. Orwant, "Heterogeneous learning in the doppelgänger user modeling system," *User Modeling and User-Adapted Interaction*, vol. 4, no. 2, pp. 107–130, 1995.

[2] G. Paliourasa, C. Papatheodoroub, V. Karkaletsisa, and C. Spyropoulosa, "Discovering user communities on the internet using unsupervised machine learning techniques," *Interacting with Computers*, vol. 12, pp. 761–791, March 2002.