qFEx - um crawler para busca e extração de questionários de pesquisa em documentos HTML

Gilney N. Mathias[1], **Carina F. Dorneles**[1]

[1]Departamento de Informática e Estatística - INE
Universidade Federal de Santa Catarina - UFSC/Florianópolis

gilney.salvo@gmail.com, carina.dorneles@ufsc.br

Abstract. *Companies or institutions can use survey questionnaires to evaluate items or products, mediate the satisfaction of their employees/customers, or be used by surveys to collect data that can be used in studies. Some problems in creating such quizzes involve: deciding which questions to ask, how to ask them, and how to organize them. With this in mind, this work proposes the creation of a* Web Crawler*, which scans the Web in search of sites that possibly contain questionnaires, and an Extractor, capable of extracting the questionnaires from the list of pages collected by the* crawler *and save them to a relational database. The database created can then serve to analyze these data and/or as a centralized base of examples to prepare new questionnaires or reuse existing questions. Some experiments were conducted to demonstrate the correct collection of questionnaires by the* crawler*, and the subsequent extraction of questions present in the questionnaires.*

Resumo. *Questionários de pesquisa podem ser utilizados por empresas ou instituições para avaliar itens ou produtos, mediar a satisfação de seus funcionários/clientes, ou serem utilizados por pesquisadores para coleta de dados que podem ser usados em estudos. Alguns problemas na criação de tais questionários envolvem: decidir quais perguntas fazer, como fazê-las e como organizá-las. Visando isso, este trabalho propõe a criação de um* Web Crawler*, que varre a Web em busca de sites que possivelmente contenham questionários, e de um Extrator, capaz de extrair os questionários da lista de páginas coletadas pelo* crawler *e salvá-las em um banco de dados relacional. A base de dados criada pode depois, servir para a análise desses dados e/ou como uma base centralizada de exemplos para a elaboração de novos questionários ou ainda para o reuso de questões existentes. Alguns experimentos são apresentados para demonstrar a correta coleta de questionários pelo* crawler*, e a posterior extração das questões presentes nos questionários.*

1. Introdução

Nos últimos anos, é indiscutível o quanto a Web trouxe facilidade na comunicação entre pessoas de vários cantos do mundo e a busca por informações e conhecimento. Tais fatos abriram as portas para o uso de questionários online, uma forma mais abrangente e fácil para empresas e pesquisadores coletarem dados ou perfis de pessoas[1]. A grande vantagem do uso de tais questionários online é a possibilidade de alcançar um grande

[1]https://www.surveymonkey.com/

número de pessoas, com características em comum e/ou únicas, de forma rápida e barata [Wright 2017].

Questionários de pesquisa normalmente são projetados pela necessidade de se obter informação cujos dados não existem – ou existem em quantidade insuficiente. Alguns problemas encontrados na fase de *design* de tais questionários incluem: decidir quais questões perguntar, qual a melhor maneira de expressá-las e como arranjar as perguntas para se obter a informação necessária. Questionários de pesquisa podem ser utilizados com várias finalidades, tais como: em empresas para avaliar produtos ou averiguar a satisfação de seus funcionários, para medir a satisfação de serviços, entre outras; em instituições de ensino e pesquisa para avaliação de seus corpos docente e discente [da Silva 2012], por exemplo; ou por pesquisadores para coleta de dados que são posteriormente usados em estudos e pesquisas. Portanto, pode ser bastante útil reusar questionários, ou parte deles, já criados para a realização de novas coletas de dados.

Algumas comunidades de pesquisa, tais como *Hirsh Health Sciences*[2], *ADAI Library*[3], *RAND Health*[4] e IHSN[5], mantêm repositórios de questionários de pesquisa que são publicamente acessíveis. Eles incluem diferentes questionários que ajudam profissionais e pesquisadores a analisar os resultados de questões, adicionar novas questões ou mesmo indicar questões sem sentido. Estes repositórios também são úteis para pesquisadores que estejam procurando por inventários de instrumentos de questionários de pesquisas já validados. Neste sentido, é interessante ter uma ferramenta que busque questionários sem a necessidade de ser um usuário registrado.

Este trabalho tem como objetivo coletar questionários de pesquisa na Web e extrair questões para construir um banco de dados centralizado das informações coletadas. Os dados extraídos servirão como uma base de conhecimento que pode ser usada como ponto de partida para a construção de novos questionários ou para a análise das características presentes visando extrair algum tipo de informação útil. Vale ressaltar que até a presente data não se encontrou nenhum trabalho que tenha esse foco na busca e extração dos dados de questionários de pesquisa na Web. São dois os principais problemas abordados por este trabalho: fazer a detecção de questionários em uma página qualquer da Web; e conseguir fazer a extração dos dados nele contidos, de forma genérica, para uma base de dados. O trabalho foi desenvolvido para realização dos experimentos relatados em [de Souza and Dorneles 2019], cuja proposta é uma métrica de similaridade para comparar questionários de pesquisa, que se caracterizam pela heterogeneidade de suas perguntas, e para fornecer um método de classificação com base nas variações de consultas construídas por um usuário em busca de questionários.

O grande desafio na identificação e extração de questionários online é a quantidade enorme de diferentes formas utilizadas em sua construção utilizando HTML [Laender et al. 2002]. Cada site tem sua própria maneira de estruturar o HTML, de estilizar os elementos da página e de enviar os dados do cliente para o servidor. No contexto de questionários de pesquisa, as páginas podem ter conteúdo estático, aonde todas as perguntas do questionário são carregadas após o usuário responder uma ou mais

[2]https://researchguides.library.tufts.edu/c.php?g=249271&p=1659301

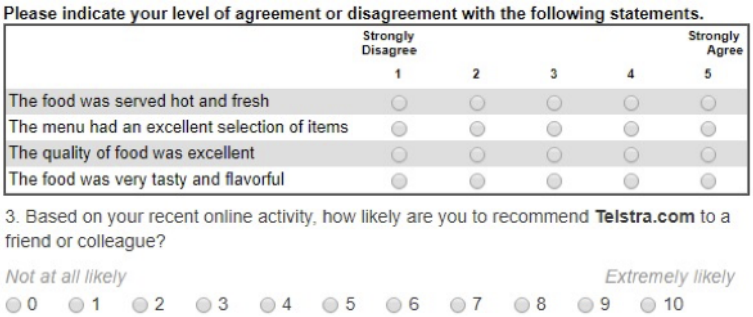
[3]http://lib.adai.washington.edu/instruments.htm

[4]https://www.rand.org/health/surveys_tools.html

[5]http://www.ihsn.org/health-modules

perguntas anteriores ou clicar em um botão de 'próximo', redirecionando o usuário para a próxima parte ou página do questionário. Desta forma, é importante ressaltar que este trabalho se foca na identificação e extração de questionários contidos em páginas com conteúdo estático. A Figura 1 apresenta uma das possíveis formas de construção de perguntas com alternativas.

Figura 1. Exemplo de uma pergunta e alternativas de respostas de um questionário de pesquisa



Para lidar com a falta de padronização na criação de questionários online utilizando HTML, propõe-se a criação de dois conjuntos de heurísticas: um para a parte de identificação de questionários em uma página HTML e outro para a extração das questões presentes nesses questionários. Além disso, é proposta a implementação de um *Web Crawler* [Olston and Najork 2010] focado, um programa que varre a Web fazendo o download de páginas que possuam certas características interessantes para o usuário [Liu 2007], bem como de um Extrator, que implementem as heurísticas mencionadas e que juntos sejam capazes de fazer a coleta dos questionários e a extração de seus dados para uma base de dados.

O artigo está organizado como segue. Na Seção 2, é apresentado o processo de funcionamento do *crawler* e do extrator propostos. Na Seção 3, é descrita a base de dados e suas principais características. Os resultados dos experimentos são apresentados na Seção 4 .Algumas aplicações e trabalhos existentes são descritos na Seção 5. Na Seção 6 são apresentadas as conclusões e as direções futuras.

2. qFEx

Esta seção apresenta o funcionamento do qFEx[6] (*Questionnaires' Finder and Extractor*). Inicialmente, é descrita uma visão geral do trabalho e em seguida são explicados os conceitos, os componentes desenvolvidos e as ferramentas utilizadas no seu desenvolvimento.

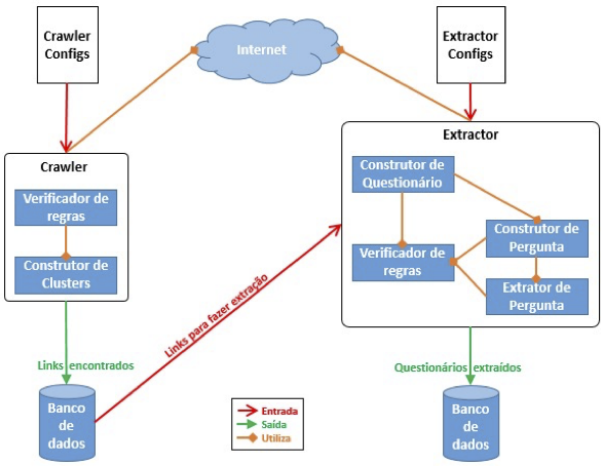
2.1. Visão Geral

A Figura 2 apresenta o processo de coleta e extração de questionários, que possui dois grandes componentes: *Crawler* e Extractor. Ambos recebem como entrada um arquivo que possui as configurações do banco de dados, do nível de *log*, da biblioteca de *crawler*, de *seeds* para busca/extração e de valores para os parâmetros utilizados. De forma geral, o processo funciona da forma tradicional: o componente *Crawler* varre a Web, utilizando

[6]https://github.com/nogenem/TCC_UFSC

como base as *seeds* fornecidas no arquivo de configuração, em busca de questionários que possuam certas características e salva os links em um arquivo; o componente Extractor, por sua vez, usa os links do *Crawler*, e o arquivo de configuração, e extrai os dados dos questionários persistindo-os de forma estruturada em uma base de dados.

Figura 2. Visão geral de funcionamento do qFEx



2.2. Detecção de Questionários

Um ponto importante a ser considerado na detecção de questionários em páginas Web foi a necessidade de encontrar características que os distinguem de outros tipos de formulários (*WebForms*) quaisquer [Santos et al. 2012]. Assim, foram definidas algumas heurísticas específicas que identificam certo componentes próprios de um questionário.

H1. Palavras Comuns. Páginas com questionários normalmente possuem certas palavras comuns em seu título e/ou em seu corpo. Algumas palavras, tais como ”questionário”e ”survey”, são bastante comuns em questionários *online* e podem ser utilizadas como uma filtragem inicial para verificar se uma dada página web possui ou não um questionário. Esta lista de palavras pode ser configurável.

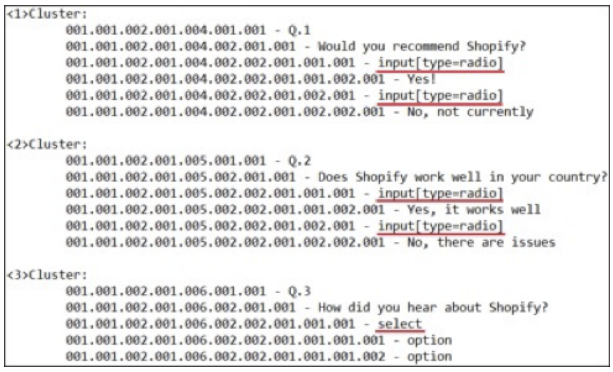
H2. *Clusters* de Componentes. É possível determinar a presença de um questionário verificando a quantidade de *clusters* (agrupamento de elementos HTML) que possuam pelo menos um componente de formulário ou que possuam uma certa quantidade de componentes de formulário. Um questionário normalmente possui uma quantidade mínima de *clusters* em sequência que possuem componentes de formulário, conforme pode ser observado no exemplo da Figura 3.

H3. Início da pergunta. A descrição de uma pergunta normalmente começa com um número ou uma palavra qualquer iniciando com letra maiúscula, possuindo pelo menos quatro caracteres, um espaço e terminando com o caractere ‘:’, ‘?’ ou ‘.’.

H4. Proximidade de elementos. Perguntas normalmente possuem suas descrições e componentes de formulário próximos uns dos outros, tais como os elementos de INPUT, TEXTAREA, SELECT e afins.

H5. Exclusão por distância. É possível eliminar certos formulários/campos que não pertencem a um questionário, verificando a distância entre os componentes e certas

Figura 3. Exemplo de *Cluster* de Perguntas com Componentes de Formulário



palavras/frases que são normalmente encontradas nos mesmos. Esta heurística é utilizada principalmente para eliminar componentes soltos pelas páginas web, como campos de busca, ou ainda para casos de formulários de login, registro e afins.

2.3. Navegação entre os nodos

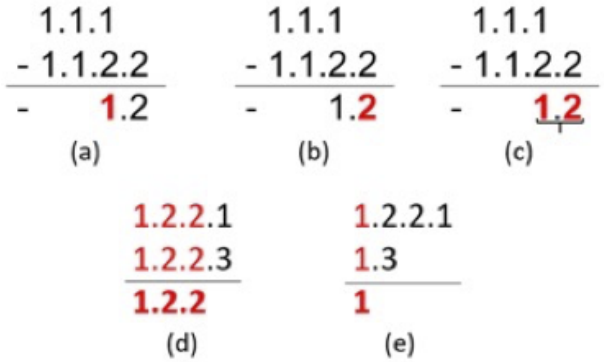
Neste trabalho, a numeração Dewey [Tatarinov et al. 2002] foi estendida e chamada de Dewey-Ext de forma a permitir o acréscimo de novos conceitos utilizados nas heurísticas empregadas. Tais conceitos servem para averiguar as relações entre os nodos, e são utilizados em quase todos os parâmetros de configuração e heurísticas do Crawler e do Extractor. Os conceitos adicionados a numeração Dewey são os seguintes:

- Height: se refere ao valor do primeiro número do ID da distância entre dois nodos. No caso da Figura 4(a), a altura é 1, ou seja, os primeiros nodos que eles têm de diferença estão em sequência na estrutura HTML;
- Max Height: é o valor do maior número do ID da distância. No exemplo da Figura 4(b), a altura máxima é 2;
- Width: é o valor do comprimento do ID da distância. No caso da Figura 4(c), este valor é 2. Isto pode ser utilizado para verificar se os nodos não estão em profundidades muito distintas na árvore;
- Common Prefix: Se refere ao maior 'prefixo comum' entre os IDs dos nodos. Este cálculo é demonstrado na Figura 4(d) e 4(e) e é feito varrendo-se os IDs da esquerda para a direita e copiando os seus valores iguais até que se encontre o primeiro valor diferente. Este resultado pode ser utilizado para verificar se três ou mais nodos possuem a mesma sequência de parentes, caso os prefixos sejam iguais, ou para se averiguar a quantidade de parentes em comum, olhando o comprimento dos prefixos.

Além dos conceitos recém introduzido, vale ressaltar outras alterações realizadas neste trabalho em relação à Numeração Dewey:

- Nodos de comentário, *tags* BR, DIV, SPAN, P, TH, e A, vazias e sem o atributo href, são ignorados na hora de montar os IDs. Isto foi feito pois esses elementos não agregam em nada para o objetivo do trabalho e poderiam atrapalhar nas medidas de distância entre os elementos do questionário;
- Nodos de texto separados por *tags* de quebra de linha, BR, são unidos em um único nodo. Isto ajuda em alguns aspectos da implementação, como a descoberta

Figura 4. Exemplo do cálculo da distância e prefixo comum entre dois nodos



dos cabeçalhos de matrizes aonde o texto dos mesmos foi quebrado com o uso da *tag* BR para uma melhor representação visual;

- Os IDs são representados com preenchimento de até dois zeros, por exemplo: o ID 1.10.100 é representado como 001.010.100. Esta forma de representação permite que algumas operações sejam facilitadas.

2.4. Implementação

O qFEx foi desenvolvido em JAVA da Oracle[7] em conjunto com as seguintes bibliotecas externas: Crawler4j[8], que possui todas as funções que um *crawler* deve implementar; Jsoup, para fazer a manipulação da árvore DOM do HTML; e Json, para lidar com arquivos no formato Json. O trabalho foi dividido em três subprojetos (*Common Lib*, *Crawler* e *Extractor*) com o intuito de diminuir a duplicidade de código.

3. Conjunto de Dados

Atualmente, os dados coletados são oriundos de 2.262 links de questionários encontrados em 32 web sites diferentes[9], distribuídos tanto em português quanto em inglês. Os questionários estão classificados em 8 domínios de pesquisa diferentes que são apresentados na Tabela 1.

Domínio	***Quantidade***	***% do Total***
Avaliação/Satisfação de/com produtos, serviços, etc	214	41,96%
Outros assuntos diversos	88	17,25%
Pesquisa de Mercado, Negócios e Marketing	76	14,90%
Recursos Humanos e Ambiente Empresarial	47	9,26%
Educação e Treinamento	30	5,88%
Saúde e Esportes	22	4,31%
Entretenimento e Eventos	22	4,31%
Comunidade e ONGs	11	2,16%

Tabela 1. Domínios de aplicação dos questionários coletados

[7]https://www.oracle.com/java/index.html
[8]https://github.com/yasserg/crawler4j
[9]https://github.com/nogenem/TCC_UFSC/blob/master/tcc_forms_v03_all.backup

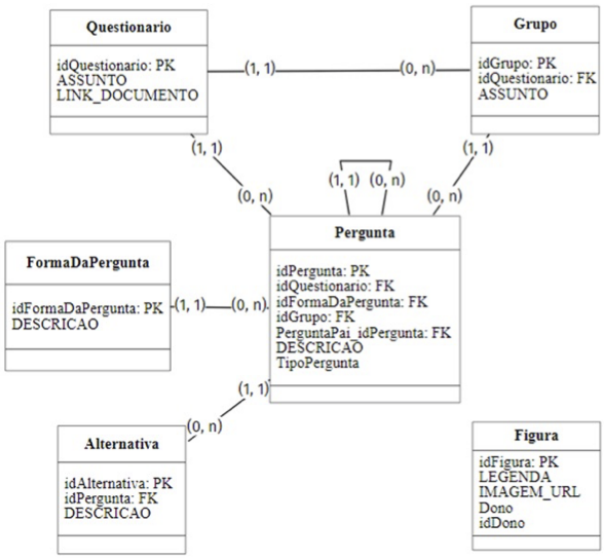
3.1. Estatísticas dos Dados coletados

Dos 2.262 links de questionários encontrados, foi necessário efetuar uma limpeza, excluindo questionários irrelevantes, tais como aqueles que possuíam menos de 2 perguntas ou que eram redundantes. Assim, a atual base de dados possui 510 questionários com 5765 perguntas no total, tendo em média 11 perguntas por questionário, sendo que o menor questionário possui 2 perguntas e o maior 53 perguntas. Deste total, 4161 são perguntas fechadas, ou seja, perguntas que possuem alternativas fixas onde o usuário tem que escolher exatamente uma delas, 1351 são de perguntas abertas, em que os usuários são livres para informar as respostas que desejarem, e 254 são de perguntas de múltipla escolha, onde os usuários podem escolher entre uma ou mais das alternativas presentes. Estes valores podem mudar, caso o *crawler* seja rodado novamente com diferentes *seeds*, ou até mesmo com as *seeds* padrão, visto que a dinâmica de atualização dos sites web é bastante variável.

3.2. Modelo de Dados

Os dados extraídos foram inicialmente estruturados em JSON, mas armazenados em um banco de dados relacional. A estrutura lógica definida foi pensada de tal forma que os questionários podem ser reconstruídos conforme os originais. A Figura 5 apresenta o modelo do banco de dados onde os dados extraídos são persistidos. A tabela `Questionario` guarda o link do questionário, como forma de manter sua origem. Alguns pontos importantes devem ser considerados: (a) a tabela `FormaDaPergunta` guarda todas as diferentes formas de se fazer uma pergunta em um questionário, tais como: CHECKBOX INPUT, RADIO INPUT MATRIX, TEXT INPUT GROUP e entre outros; (ii) a tabela `Figura` guarda as informações de imagens encontradas no questionário; tais imagens podem pertencer ao questionário, a uma pergunta ou a uma alternativa, campos `Dono` e `idDono`; (iii) uma pergunta pode ter uma ou mais perguntas filhas. Isto é usado para casos como matrizes e perguntas com subperguntas; e (iv) uma pergunta pode ser do tipo ABERTO, FECHADO ou MULTIPLA ESCOLHA, dependendo da sua forma.

Figura 5. Modelo do Banco de Dados



4. Avaliação e Resultados

Nesta seção, são apresentados resultados de revocação, precisão e f-value encontrados com os experimentos realizados com o *Crawler* e o Extractor. O objetivo com estes experimentos foi verificar se o *Crawler* identificou corretamente as páginas HTML que possuíam questionários de pesquisa. Além disso, foi avaliada a correta extração de perguntas, alternativas de cada pergunta e extração de perguntas filhas.

A precisão do *Crawler* atingiu 94,47%, ou seja, 2137 dos links encontrados pelo *Crawler* realmente possuíam um ou mais questionários neles. Notou-se que outros 5,53% dos links encontrados não possuíam questionários, mas sim formulários muito grandes ou múltiplos formulários pequenos, porém próximos uns dos outros, o que acabou confundindo o algoritmo de detecção de questionários do *Crawler*. No caso do *Crawler*, não foi realizada avaliação de revocação, visto que a coleta foi realizada em toda a Web.

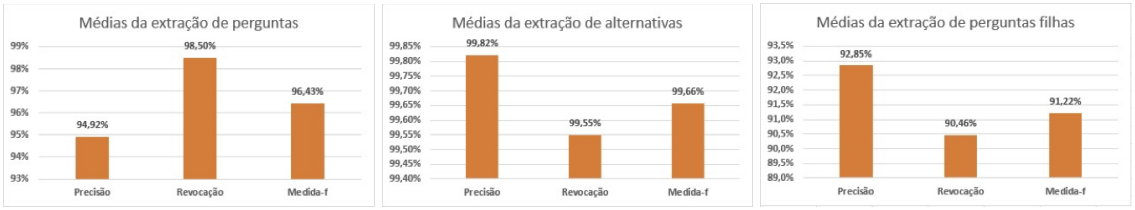


Figura 6. Revocação, Precisão e F-Value para (a) Perguntas; (b) Perguntas-Filhas; (c) Respostas

A Figura 6(a) contém as médias gerais das métricas utilizadas para avaliar a extração de perguntas. Como pode ser visto, o resultado das três métricas ficaram acima de 90%, o que indica que a maioria das descrições de perguntas foram extraídas corretamente. A Figura 6(b) apresenta as médias gerais da extração de alternativas. Neste caso, percebe-se que os resultados ficaram bem próximas de 100%, o que indica que a abordagem usada para a extração das alternativas foi eficaz.

A Figura 6(c) exibe os resultados para a extração de perguntas filhas, aonde é possível ver que a precisão chegou a um pouco mais que 92%, a revocação a 90% e a medida-f a 91%. Existem dois motivos principais para estes resultados reduzidos em comparação a extração de perguntas e alternativas: (i) alguns sites colocam as perguntas filhas no mesmo nível que uma pergunta normal, fazendo com que o Extractor se confunda e as considere como perguntas normais; e (ii) uma deficiência no reconhecimento da descrição da pergunta pai de uma pergunta com subperguntas faz com que, em alguns casos, o assunto do questionário seja considerado como pergunta pai da(s) primeira(s) pergunta(s) do questionário. Isto ocorre porque, nestes casos específicos, a estrutura do início do questionário fica idêntica a estrutura de uma pergunta com subpergunta.

5. Aplicações e Trabalhos Existentes

Na literatura é possível encontrar trabalhos relevantes para diferentes comunidades (e.g. Web Semântica, Saúde, Educação, Ciência Aberta, entre outras) que fazem uso de dados de questionários para avaliação de resultados, ou realização de tarefas específicas. Nesta seção, alguns desses trabalhos são agrupados por tarefas, com o objetivo de corroborar a relevância dos dados de questionários coletados.

Knowledge base augmentation. Enriquecimento de bases de conhecimento, também conhecido como ”população de base de conhecimento” [Zhang and Balog 2019], preocupa-se em gerar novas instâncias de relações usando dados tabulares e atualizando as bases de conhecimento com as informações extraídas.

Avaliação de uso de recursos. Questionários de pesquisa fornecem subsídios importantes para análise e avaliação de recursos utilizados em pesquisas científica. Por exemplo, no domínio de saúde[10], tem como objetivo ajudar médicos e pesquisadores a encontrar instrumentos usados para triagem e avaliação do uso de substâncias e transtornos por uso de substâncias. Esses questionários são elaborados para uma ampla gama de finalidades, incluindo avaliação da saúde dos pacientes, rastreamento de condições de saúde mental e medição da qualidade do atendimento e da qualidade de vida.

Reprodutibilidade de resultados científicos. Muitas vezes, para que exista fidelidade na reprodutibilidade de um resultado científico, é necessário que metologia seguida seja a mesma, incluindo os materiais e métodos usados [Lobo et al. 2008]. Desta forma, é de suma importância prover a disponibilização online e pública dos questionários de pesquisa utilizados.

Avaliação de questionários utilizados em pesquisas. As respostas fornecidas a essas perguntas servem como base para análises científicas e fornecem a base de indicadores estatísticos usados para descrever o estado de uma sociedade. Obviamente, esses dados são tão significativos quanto as perguntas que fazemos e as respostas que recebemos. Além disso, a quem perguntamos é de importância crucial para nossa capacidade de tirar conclusões que vão além das pessoas específicas que responderam às nossas perguntas. Consequentemente, os processos subjacentes às respostas às perguntas e à seleção adequada dos entrevistados são de grande importância para muitas áreas da pesquisa social [Schwarz 2007].

Template para novos questionários. A reutilização de questionários existentes pode ser de grande ajuda no *design* e desenvolvimento de novos questionários. Questionários de sucesso, aplicados em pequenas amostras, podem ser reproduzidos em outras populações com custo mais baixo de desenvolvimento.

6. Conclusões e Trabalhos futuros

Os principais problemas abordados por este trabalho foram a descoberta e extração de questionários de pesquisa encontrados na Web. Para resolver tais problemas foram desenvolvidos um Web Crawler focado capaz de varrer a Web em busca de questionários online e um Extrator que consegue extrair os dados dos questionários para um banco de dados relacional. Como trabalhos futuros, pode-se apontar algumas direções interessantes: (i) adicionar suporte à detecção e extração de questionários em páginas com conteúdo dinâmico; (ii) melhorar a extração de perguntas filhas; (iii) implementar um algoritmo que consiga detectar o idioma em que os questionários estão escritos (português, inglês, espanhol, etc); e (iv) análise do uso de perguntas idênticas em questionários distintos. Isto pode ser útil para a análise dos dados.

[10]https://www.rand.org/health-care/surveys_tools.html

Referências

da Silva, J. M. (2012). Collecta: um sistema computacional de coleta de dados e avaliação institucional para apoio à tomada de decisão na universidade federal de santa catarina. Master's thesis, Universidade Federal de Santa Catarina, Florianópolis.

de Souza, R. H. and Dorneles, C. F. (2019). Searching and ranking questionnaires: An approach to calculate similarity between questionnaires. In *Proceedings of the ACM Symposium on Document Engineering 2019*, DocEng '19, New York, NY, USA. Association for Computing Machinery.

Laender, A. H., Ribeiro-Neto, B., da Silva, A., and Teixeira, J. (2002). A Brief Survey of Web Data Extraction Tools. *Sigmod Record*, 31(2).

Liu, B. (2007). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-Centric Systems and Applications. Springer.

Lobo, A. S., de Assis, M. A. A., de Barros, M. V. G., Calvo, M. C. M., and Freitas, S. F. T. (2008). Reprodutibilidade de um questionário de consumo alimentar para crianças em idade escolar. *Revista Brasileira de Saúde Materno Infantil*, 8(1).

Olston, C. and Najork, M. (2010). Web Crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246.

Santos, L., Dorneles, C. F., and Mello, R. d. S. (2012). An approach for extracting web form labels based on distance analysis of html components. In *IADIS WWW/Internet Conference*.

Schwarz, N. (2007). *Evaluating Surveys and Questionnaires*, pages 54–74.

Tatarinov, I., Viglas, S. D., Beyer, K., Shanmugasundaram, J., Shekita, E., and Zhang, C. (2002). Storing and querying ordered xml using a relational database system. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, SIGMOD '02, page 204–215, New York, NY, USA. Association for Computing Machinery.

Wright, K. B. (2017). Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, 10(3). JCMC1034.

Zhang, S. and Balog, K. (2019). Knowledge base augmentation. In *SIGIR 2019 tutorial, https://iai-group.github.io/webtables-tutorial/slides/part-3.pdf*, New York, NY. ACM.