# Assignment 3

Author: Matthias Tilsner
Student-ID: 200882645
Course: COMP 6785

Submission: February 4, 2009

## 1 From the two topics selected in Assignment 2, choose one topic as your primary research topic. This topic will form the basis for the remaining assignment, presentation, and project. Provide a detailed description of the problem domain, the data to be visualized, and the characteristics of the target users.

The topic I choose is “information retrieval on the web”, thus concentrating on web based search engines.

Web search engines are used for retrieving information from the web. They expect a *query string* that consists of a number of *search terms*. Each term is a single word or a fragment of a sentence enclosed in quotes. *Search terms* can be ambiguous. “Java” for example is both an island and a programming language. “Jobs” might represent vocations or the current CEO of Apple Industries. A search engine processes the *query string* by comparing its *search terms* with a set of available documents. For web searches, this set will be all web pages accessible by the search engine. The search engine compares the content of the documents with the different *search terms*, trying to find out the applicability of the individual documents. The results are typically returned as an ordered list. This order is the main benefit of a classic search engine, since it is the only semantic information generated. The idea is that the relative position of a a document in the list corresponds to the chance that this paper contains what the user is looking for. The position of a document is also called its “ranking”.

The exact algorithm used for calculating the ranking of a document is kept secret by the search engine providers. It is however known that it is derived from multiple sources: The frequency that a *search term* occurs inside a document displays how relevant the document might be in regard to that term. If multiple *search terms* are used, the document with the highest combined frequency is likely to be the one most relevant to the full search query. Furthermore, web documents are connected to one another using *links*. Most current search engines utilize these links in order to deduct the quality of a document. The more links target it, the more people suggest it, the better the quality. The quality is taken into consideration when the ranking of a document is evaluated.

Documents with a high ranking will be the first documents to be displayed, while documents with a low ranking will typically not even be visited by users. Because of that, web pages commonly court for better rankings. In order to achieve this, many pages try to falsely increase the numbers of links targeting itself and the number of occurrences of certain common *search terms*. This causes a considerable loss of search quality, reduces the effectiveness of a search and makes it harder for the user to actually find what he is looking for.

Web documents typically consist of both textual and graphical information. For search engines, only the textual information is relevant, since web development paradigms specifically require textual meta-data in addition to every graphic used. This is necessary to provide accessibility for people with disabilities. Current helper-software is not able to process these graphics. Screen-readers, for example, rely on additionally provided information. It can therefore be assumed, that critical content encoded in graphics is always redundantly provided as meta-data.

Links cannot only be used for identifying the quality of a document, but also for identifying how different documents are connected. By taking the links into account when constructing the search result, relationships can be extrapolated. These relationships often indicate a semantical similarity, thus allowing a categorization of the documents. Unfortunately, most current web search engines do not use this option efficiently. Also, some documents might actually be sub-documents of one another. For example, a web site might have several pages. The documents are not only related to one another, but are actually different elements of one single document. It might furthermore be, that the different sub-documents actually refer to the same document. This relationship is usually displayed by web browsers.

Web search engines are used by a broad variety of people, ranging from professionals to novice users. It is important to support both kind of users, both the ones with much experience in searching and browsing the web, and those who have no idea how search engines, the Internet, or maybe even computers themselves are operated. Professional users might require a more flexible and powerful search interface, providing multiple features in order to narrow down the search results in advance. Non-professionals, however, have to be able to understand and use the interface intuitively. Both users have a few things in common: searching for a document is not their main task. Reading and processing the document is. Therefore, the search process has to run as quick and smooth as possible. Consequently, users will not be willing to spend much time learning the interface or adapting to changes. Search engines must therefore be as intuitive as possible. Also, users can be expected to be impatient with search engine, not wanting to spend too much time on the search task. Therefore, search engines must operate efficiently with great performance. Performing complex calculations on runtime might compromise this performance.

# 2 Perform an in-depth literature review of this topic that covers both the seminal work, as well as recent advancements. For each paper you find that is relevant to the topic, provide a brief summary of its contribution to the field.

The following papers are also relevant. Unfortunately, I did not have the time to summarize them: [Andrews et al., 2001, Berenci et al., 1999, Brin and Page, 1998, Chang and Hsu, 1999, Chirita et al., 2005, Cutting et al., 1992, Eick, 1994, Hearst, 1995, Heimonen and Jhaveri, 2005, Jansen and Pooch, 2001, Joho et al., 2004, Jones, 1998, Kobayashi and Takeda, 2000, Kosala and Blockeel, 2000, Kules et al., 2006, Paek et al., 2004, Radlinski and Dumais, 2006, Robertson and Jones, 1997, Sanderson and Croft, 1999, Spink, 2002, Spink et al., 2001, Sugiyama et al., 2004, Teevan et al., 2005, Vaughan, 2004, Wise et al., 1995, Zamir and Etzioni, 1998]

## 2.1 [Alonso and Baeza-Yates, 2000]

This paper discusses the dependability of information retrieval components and the InfoVis client. Its main concern is the visualization of large document collections in web search engines. It proposes an architecture, that allows the reuse of visualization results, thus making them accessible to multiple users. This is extremely helpful in collaboration applications and "solves the problem on data overload on Internet".

## 2.2 [Alonso and Baeza-Yates, 2003]

The paper tries to increase the comprehensibility of web search results by proposing a split view interface. This interface displays both the results in an ordered manner, and a short overview over selected search results, highlighting the individual search terms so that the reader can get a grasp of the relevance of the document more easily. While developing the solution, the authors concentrated primarily on performance and scalability. This is achieved by letting the server perform all visualization, thus reducing the client effort.

## 2.3 [Benjamin et al., 2008]

This paper introduces a method to visualize similarities in web results, thus providing a way for detecting plagiarism. The presented solution uses Kohonen Maps to visualize the similarities, aiming for the user to be able to grasp it intuitively. It furthermore incorporates a Customized Term Weighting Scheme that extracts semantic information, thus enabling an automated classification of the search results.

## 2.4 [Einsfeld et al., 2006]

This paper aims to provide a method for visualizing both content and semantic information of documents. This aim is constricted by the objective of providing an intuitive user interface. It proposes an application called *DocuWorld* that visualizes documents content and meta-data together with the semantic relations between documents in a 3 dimensional canvas. It includes a method for adapting the visualization based on user input and the currently visible context.

## 2.5 [Hoeber, 2007]

This thesis concentrates on the interactive refinement of web search terms inside existing visualizations and the exploration of results. It proposes four different tools for achieving those tasks: HotMap allows the user to add and remove search terms while exploring the result list. Furthermore, the user can define the importance of the search terms, thus allowing an interactive reordering of the result list. While HotMap is based on classic one-dimensional result lists, VisiQ provides a two dimensional displayal of the results allowing the user to easily comprehend how documents are related to one another. WordBar displays the terms most commonly used inside the documents of the result list. This list is very helpful for identifying additional terms that could be added to the query for refining the result list or that should be excluded (i.e. by using googles "-" prefix). The last tool presented is the Concept Clusterer, clustering documents by using color for encoding. Two user studies were conducted, showing positive results especially for the WordBar and the HotMap tool. Even though the Concept Clusterer was helpful an additional step was required including reevaluation of the results. VisiQ was not used in the studies.

The thesis greatly proposes the utilization of interactive tools in order to enable the user to iteratively refine an specify his search. Furthermore, it suggest a two-dimensional design space plotting the user tasks in relation to generated information about the result items at hand. In order to allow a visual representation of the search terms, a Concept Knowledge Base is introduced. Next to the benefits generated by the tools, the thesis also presents a novel evaluation method for specifically evaluating InfoVis applications for web searches.

## 2.6 [Konchady et al., 1998]

This paper tries to provide a new way for visualizing search results. It concentrates on the relevance of the search results in correspondence to the query words. It displays all search results in a multi-dimensional environment (up to 3 dimensions). The user can assign the different search terms or documents to the axes. All other documents are displayed by dots with their positioning being derived from the similarity of the document with the axis value. The user can then rotate the cube and zoom in and out in order to retrieve information. Furthermore, the paper proposes a link analysis tool that extrapolates the relationship of documents between one another by looking at the links, and one visualizing the similarities of groups of people regarding their search terms.

## 2.7 [Weiss-Lijn et al., 2001]

This paper aims to provide an efficient way for querying documents. It uses meta-data that describes and differentiates the content of different paragraphs, a document consists of. Instead of the original document content, only the meta-data is visualized. A test experiment was run that inspired redesign the the application in order to improve the performance. A numerical advantage could not be proven. However, the paper suggests that this is due to utilization of meta-data.

### 2.8 [Mukherjea and Hara, 1999]

This paper proposes an alternative interface for displaying the results of a web search engine, providing the user with the means to navigate through the documents based upon their relationship. It uses sets of cards that are layered in front of each other to display related documents via connecting lines. The relevance is displayed by assigning different colors to the search terms and coloring the individual result documents accordingly.

### 2.9 [Mukherjea et al., 1996]

This paper proposes the visualization of search results using either a scatter point matrix, or displaying the search results in a three dimensional space. Color and dimensions of the individual document points inside the graphics are used for encoding the relevance to the search terms. No usability study is performed within the paper. Instead, one is proposed, suggesting the refinement of the findings according to the results of that study.

### 2.10 [Nowell et al., 1996]

This paper discusses the visualization of library items with a tool called *Envision*. Here, search result items are displayed using icons that are positioned in a matrix and using different shapes and colors for the icons. The meaning of the different means of visualization can be defined by the user. The paper discusses the results of a usability test performed on *Envision*, showing positive results and user feedback. Even though the paper solely discusses the utilization for the content of a predefined library, I believe that the results are just as relevant for the area of visualizing web search results.

### 2.11 [Paulovich et al., 2008]

This paper presents the *Project explorer for the WWW*, an application that visualizes the results of web search engines in a multidimensional map. The positioning of the the different result items is dependent on their content. Similar documents are placed near to one another, thus allowing a preemptive processing of related document groups. Additionally, they are connected with lines, thus allowing an easy identification of related documents. The relevance to the groups specific search terms is displayed by coloring the search result items.

### 2.12 [Rauber and Bina, 2000]

This paper suggests a redesign of web search engine interfaces oriented on classical libraries. It proposes that the results are organized by topics and displayed in a bookshelf-like manner. The meta-data of the individual search result items is to be displayed so that it can be comprehended intuitively. The language of a document found, for example, is encoded using color (i.e. German = yellow, English = blue). The actual size of the returned document is displayed by the size of the icon representing it in the library view. Since the additional effort required for generating the visualization is minimum, the authors assume the result to be effective.

### 2.13 [Roussinov, 1999]

This paper evaluates the efficiency of making result maps used for displaying Internet search results interactive. These result maps group results together based on similarity and search term relevance. The author tries to enhance the search experience by allowing him to respecify the search terms while viewing the visualization. Furthermore, the system can generalize the grouping, thus creating fewer groups inside the result map. In a case study, this interactive features proved to be successful and were repetitively used by the users.

### 2.14 [Sebrechts et al., 1999]

This paper evaluates the effectiveness of using visualization tools per se for visualizing search results. It runs a series of test with a mixed group of both professional and non-professional users, using an InfoVis tool called *NIRVE*. The results of these tests are, that the best way for displaying the the information is dependent on the task at hand. While looking for a specific target is much easier when using a one-dimensional, textual visualization, clustering and categorization benefits greatly from multidimensional visualizations.

### 2.15 [Tvarozek and Bielikova, 2008]

This paper proposes web search results to be displayed using of hierarchical clusters and assigning them attributes with meta-data. It suggests, that the visual presentation should also include the qualitative relationship of documents. Using positioning, connections, and coloring, the paper proposes a two dimensional visualization of the search results. In this visualization, only the most basic information should be viewed at first, enabling the user to interact with the visualization in order to explore.

### 2.16 [Weippl, 2001]

This paper analyzes different methods for visualizing collections of hypertext, grouping them into clusters. It discusses two different two-dimensional user interfaces and one three-dimensional one. It proposes the use of a Self-Organized Map but at the same time identifies it as a bottleneck when processing high numbers of hypertexts. The authors announce further research in this area, trying to solve this issue.

### 2.17 [Zaina and Baranauskas, 2005]

This paper propose a two-dimensional graph displaying the results of a web search connected with lines to display relationships. The relevance of the item in regard to the search term is encoded using the result item's icon size. The semantic information can be retrieved by fetching the first 50kB for each search result. A table view displays additional non-semantic information. A case study is performed proving the solution's success to be moderate. A significant amount of the users actually preferred the table view, corresponding to a classical search result list.

# 3 Briefly summarize how the type of data affects the set techniques available for generating an effective visual representation of the data.

The type of the data that is to be visualized is a collection of plain text documents with links among each other. Therefore, the following points of interest arise. Each point of interest represents one attribute of the information to display:

1. How relevant is a document in respect to a specific search term
2. How relevant is a document in respect to all search terms
3. How are documents related to one another?
4. What topics does a document cover

Both 1 and 2 are ordinal attributes. [Bertin, 1967] suggests using size, brightness (here the brightness is called “value”), and texture for encoding them. [Mackinlay, 1986] furthermore proposes position, saturation, and color hue. Attributes 3 and 4 are nominal, representing an association and a selection respectively (brightness is here refered to as “density”). While [Bertin, 1967] proposes texture, color, shape, and orientation for encoding of attribute 3 and size, texture, value, and color for 4, position, color, texture, containment, and connection are listed in [Mackinlay, 1986] for both. Length, slope, and angle are never proposed for encoding information. Furthermore, no quantitative encoding type is required. Table 1 summarizes these findings. Obviously, there are different ways for encoding each attribute. Furthermore, most encodings are ambiguous. Consequently, no encoding-attribute association is impelled. All associations can be freely determined.

| | Attribute 1 | Attribute 2 | Attribute 3 | Attribute 4 |
|---|---|---|---|---|
| position | x | x | x | x |
| length | | | | |
| angle | | | | |
| slope | | | | |
| area / volume | x | x | | x |
| orientation | | | x | |
| brightness | x | x | | |
| saturation | x | x | | |
| color hue | x | x | x | x |
| texture | x | x | x | x |
| connection | | | x | x |
| containment | | | x | x |
| shape | | | x | |

Table 1: Possible visualization encodings for the four different attributes

# References

[Alonso and Baeza-Yates, 2000] Alonso, O. and Baeza-Yates, R. (2000). A model and software architecture for search results visualization on the www. In *Proceedings of the Seventh International*

*Symposium on String Processing Information Retrieval*, page 8, Washington, DC, USA. IEEE Computer Society.

[Alonso and Baeza-Yates, 2003] Alonso, O. and Baeza-Yates, R. (2003). Alternative implementation techniques for web text visualization. In *Proceedings of the First Latin American Web Congress*, pages 202–204.

[Andrews et al., 2001] Andrews, K., Gütl, C., Moser, J., Sabol, V., and Lackner, W. (2001). Search result visualisation with xfind. In *Proceedings of the Second International Workshop on User Interfaces to Data Intensive Systems*, pages 50–58.

[Benjamin et al., 2008] Benjamin, C., Woon, W., and Wong, K. (2008). A graphical and convenient tool for document comparison and visualization. In *Proceedings of the International Conference on Computer and Communication Engineering*, pages 362–367.

[Berenci et al., 1999] Berenci, E., Carpineto, C., Giannini, V., and Mizzaro, S. (1999). Effectiveness of keyword-based display and selection of retrieval results for interactive searches. In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, pages 106–125, London, UK. Springer-Verlag.

[Bertin, 1967] Bertin, J. (1967). *Semiologie Graphique: Les Diagrammers, Les Reseaux, Les Cartes.* Editions Gauthier-Villars, Paris.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V.

[Chang and Hsu, 1999] Chang, C.-H. and Hsu, C.-C. (1999). Enabling concept-based relevance feedback for information retrieval on the www. *IEEE Transactions on Knowledge and Data Engineering*, 11(4):595–609.

[Chirita et al., 2005] Chirita, P. A., Nejdl, W., Paiu, R., and Kohlschütter, C. (2005). Using odp metadata to personalize search. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185, New York, NY, USA. ACM.

[Cutting et al., 1992] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, New York, NY, USA. ACM.

[Eick, 1994] Eick, S. G. (1994). Graphically displaying text. *Journal of Computational and Graphical Statistics*, 3:127–142.

[Einsfeld et al., 2006] Einsfeld, K., Agne, S., Deller, M., Ebert, A., Klein, B., and Reuschling, C. (2006). Dynamic visualization and navigation of semantic virtual environments. In *Proceedings of the Tenth International Conference on Information Visualization*, pages 569–574.

[Hearst, 1995] Hearst, M. A. (1995). Tilebars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 59–66, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

[Heimonen and Jhaveri, 2005] Heimonen, T. and Jhaveri, N. (2005). Visualizing query occurrence in search result lists. *Proceedings of the Ninth International Conference on Information Visualisation*, pages 877–882.

[Hoeber, 2007] Hoeber, O. (2007). *A Study on Interactive Visualization for Web Information Retrieval.* PhD thesis, University of Regina.

[Jansen and Pooch, 2001] Jansen, B. J. and Pooch, U. (2001). A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52(3):235–246.

[Joho et al., 2004] Joho, H., Sanderson, M., and Beaulieu, M. (2004). A study of user interaction with a concept-based interactive query expansion support tool. In *Proceedings of the 26th European Conference on Information Retrieval Research*, pages 42–56. Springer Verlag.

[Jones, 1998] Jones, S. (1998). Graphical query specification and dynamic result previews for a digital library. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, pages 143–151, New York, NY, USA. ACM.

[Kobayashi and Takeda, 2000] Kobayashi, M. and Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys*, 32(2):144–173.

[Konchady et al., 1998] Konchady, M., D'Amore, R., and Valley, G. (1998). A web based visualization for documents. In *Proceedings of the 1998 Workshop on New Paradigms in Information Visualization and Manipulation*, pages 13–19, New York, NY, USA. ACM.

[Kosala and Blockeel, 2000] Kosala, R. and Blockeel, H. (2000). Web mining research: A survey. *ACM SIGKDD Explorations Newsletter*, 2(1):1–15.

[Kules et al., 2006] Kules, B., Kustanowitz, J., and Shneiderman, B. (2006). Categorizing web search results into meaningful and stable categories using fast-feature techniques. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 210–219, New York, NY, USA. ACM.

[Mackinlay, 1986] Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141.

[Mukherjea and Hara, 1999] Mukherjea, S. and Hara, Y. (1999). Visualizing world-wide web search engine results. In *Proceedings of the 1999 IEEE International Conference on Information Visualization*, pages 400–405.

[Mukherjea et al., 1996] Mukherjea, S., Hirata, K., and Hara, Y. (1996). Visualizing the results of multimedia web search engines. In *Proceedings of the 1996 IEEE Symposium on Information Visualization*, pages 64–65, 122.

[Nowell et al., 1996] Nowell, L. T., France, R. K., Hix, D., Heath, L. S., and Fox, E. A. (1996). Visualizing search results: some alternatives to query-document similarity. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 67–75, New York, NY, USA. ACM.

[Paek et al., 2004] Paek, T., Dumais, S., and Logan, R. (2004). Wavelens: A new view onto internet search results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 727–734, New York, NY, USA. ACM.

[Paulovich et al., 2008] Paulovich, F., Pinho, R., Botha, C., Heijs, A., and Minghim, R. (2008). Pex-web: Content-based visualization of web search results. pages 208–214.

[Radlinski and Dumais, 2006] Radlinski, F. and Dumais, S. (2006). Improving personalized web search using result diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 691–692, New York, NY, USA. ACM.

[Rauber and Bina, 2000] Rauber, A. and Bina, H. (2000). "'andreas, rauber'? conference pages are over there, german documents on the lower left...": an "old-fashioned" approach to web search results visualization. In *Proceedings of the 11th International Workshop on Database and Expert Systems Applications*, pages 615–619.

[Robertson and Jones, 1997] Robertson, S. E. and Jones, K. S. (1997). Simple proven approaches to text retrieval. Technical Report TR356, Cambridge University Computer Laboratory.

[Roussinov, 1999] Roussinov, D. (1999). Internet search using adaptive visualization. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, pages 69–70, New York, NY, USA. ACM.

[Sanderson and Croft, 1999] Sanderson, M. and Croft, B. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213, New York, NY, USA. ACM.

[Sebrechts et al., 1999] Sebrechts, M. M., Cugini, J. V., Laskowski, S. J., Vasilakis, J., and Miller, M. S. (1999). Visualization of search results: a comparative evaluation of text, 2d, and 3d interfaces. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–10, New York, NY, USA. ACM.

[Spink, 2002] Spink, A. (2002). A user-centered approach to evaluating human interaction with web search engines: an exploratory study. *Information Processing and Management: an International Journal*, 38(3):401–426.

[Spink et al., 2001] Spink, A., Wolfram, D., Jansen, M. B. J., and Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234.

[Sugiyama et al., 2004] Sugiyama, K., Hatano, K., and Yoshikawa, M. (2004). Adaptive web search based on user profile constructed without effor from users. In *Proceedings of the 13th International Conference on World Wide Web*, pages 675–684, New York, NY, USA. ACM.

[Teevan et al., 2005] Teevan, J., Dumais, S. T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 449–456, New York, NY, USA. ACM.

[Tvarozek and Bielikova, 2008] Tvarozek, M. and Bielikova, M. (2008). Personalized view-based search and visualization as a means for deep/semantic web data access. In *Proceeding of the 17th International Conference on World Wide Web*, pages 1023–1024, New York, NY, USA. ACM.

[Vaughan, 2004] Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing and Management: an International Journal*, 40(4):677–691.

[Weippl, 2001] Weippl, E. (2001). Visualizing content based relations in texts. In *Proceedings of the Second Australasian User Interface Conference*, pages 34–41.

[Weiss-Lijn et al., 2001] Weiss-Lijn, M., McDonnell, J., and James, L. (2001). Visualising document content with metadata to facilitate goal-directed search. In *Proceedings of the Fifth International Conference on Information Visualisation*, pages 71–76.

[Wise et al., 1995] Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. (1995). Visualizing the non-visual: spatial analysis and interaction with information from text documents. *Proceedings of the 1995 IEEE Symposium on Information Visualization*, pages 51–58.

[Zaina and Baranauskas, 2005] Zaina, C. M. and Baranauskas, M. C. C. (2005). Revealing relationships in search engine results. In *Proceedings of the 2005 Latin American conference on Human-computer Interaction*, pages 120–127, New York, NY, USA. ACM.

[Zamir and Etzioni, 1998] Zamir, O. and Etzioni, O. (1998). Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 46–54, New York, NY, USA. ACM.