# Die Analyse heterogener Unternehmensdatenbestände als Basis für die Visualisierung von Relationen in Suchergebnismengen*

*Sonja Öttl, Sonja Hierl, Bernard Bekavac & Josef Herget*

SII – Swiss Institute for Information Research
Hochschule für Technik und Wirtschaft (HTW) Chur
Ringstrasse/Pulvermühlestrasse 57
CH-7004 Chur, Schweiz
*{sonja.oettl, sonja.hierl, bernard.bekavac, josef.herget}@fh-htwchur.ch*

**Abstract**
Das Projekt „Visual Relations" verfolgt das Ziel, die Suche in heterogenen Datenbeständen visuell zu unterstützen und Relationen innerhalb der Treffermengen aufzuzeigen. Der Anwender soll hierbei durch geeignete Visualisierungen unterstützt werden, um immanente Strukturen und Verbindungen leichter erkennen zu können. Hierzu soll vor allem die Anzeige von geographischen und zeitlichen Bezügen untersucht werden.

## 1 Projektkontext

Mit der zunehmenden Digitalisierung von Inhalten aller Arten und der Zunahme von kollaborativen Anwendungen geht ein stetiges Wachstum der jährlich anfallenden Datenmengen in Unternehmen einher, wobei nur ungefähr 20% der gesamten Unternehmensdaten effektiv zur Wertschöpfung genutzt werden [Dragoon 03] und ein durchschnittlicher Arbeitnehmer mit Bürotätigkeiten zwischen 15% und 35% seiner Arbeitszeit bei der Suche nach Informationen verbringt [Feldman 04].

Das Projekt Visual Relations greift diese Problematik auf durch die Entwicklung einer geeigneten Visualisierung zur Darstellung heterogener Datenmengen sowie

* Veröffentlicht in: Osswald, Achim; Stempfhuber, Maximilian; Wolff, Christian (Hrsg.) (2007). Open Innovation. Proc. 10. Internationales Symposium für Informationswissenschaft. Konstanz: UVK, 333-341.

inhärenter Zusammenhänge und Korrelationen von Suchergebnissen unter Berücksichtigung von zeitlichen und geographischen Bezügen.

## 2 Entwicklung von Datenmengen in Unternehmen

Betrachtet man die Entwicklung bei Unternehmensdaten, die kontinuierlich gespeichert bzw. verwaltet werden, so ist ein stetiges Wachstum zu beobachten. Einer Studie von Forrester Research nach, lagen 2006 weltweit bereits rund 318 Petabyte an Unternehmensdaten vor [Balaouras et al. 2006]. Für die kommenden fünf Jahre wird ein jährliches Wachstum um rund 20%-30% prognostiziert, das bis im Jahr 2011 zu einer Datenmenge von rund 3243 Petabyte führt.

## 3 Herausforderung bei der Integration einer Geografischen Informationssystem-Komponente in die Ergebnisvisualisierung

Auf dem Markt gibt es bereits einige Suchsysteme, die eine Visualisierung der Ergebnismenge vornehmen, wobei die Dokumentenrepräsentation im Vordergrund steht. In der Regel wird dabei eine Visualisierung der potenziellen thematischen Zusammengehörigkeit von Dokumenten vorgenommen, beispielsweise durch Clustering (zum Beispiel bei Grokker) oder durch den Einsatz von topographischen Metaphern (zum Beispiel bei Kartoo).

Die wenigsten Systeme dieser Art ermöglichen jedoch die visuelle Aufbereitung von zeitlichen oder geographischen Bezügen. Gerade diese Aspekte sind in diversen Anwendungsgebieten, beispielsweise in der Kriminalitätsbekämpfung oder im Bereich der Logistik, jedoch von höchster Relevanz. Ein Grund für fehlende Systemlösungen dieser Art mag in der Komplexität und den Herausforderungen bei der Umsetzung von Zeit- und GIS-Komponenten in Suchsystemen mit Visualisierungskomponenten liegen. In wie weit die Bezüge und Relationen innerhalb von Trefferdokumenten vorhanden sind, wurde zunächst anhand von Testkollektionen untersucht.

## 4 Analyse von Testkollektionen

Im Rahmen des Projektes wurden vier Testkollektionen untersucht, die bezüglich ihres Homogenitätsgrades sowie ihres Umfangs variieren. Die Untersuchung der

Testkollektionen erfolgte in mehreren Schritten. Nach einer Gesamtsichtung der Dokumente wurden maschinell sämtliche „File System Object" (FSO)-Attribute ausgelesen und anschließend mittels der Software „R"[1] ausgewertet und visualisiert. Dabei wurden die einzelnen Attribute auf allfällige Korrelationen untersucht, die Zusammensetzung der Kollektionen hinsichtlich Größe, Dateityp etc. fixiert und – sofern sinnvoll – statistische Mittelwerte gebildet.

Im Anschluss wurden die einzelnen Testkollektionen intellektuell erschlossen. Hierzu wurden Stichproben per Zufallsprinzip erhoben, die anschließend gesichtet wurden. Die als zentral erachteten Termini wurden zu jeder Stichprobe ermittelt und in einer Begriffsmatrix notiert. Das Ziel hierbei war, ein Relationsschema zu fixieren, anhand dessen das zu entwickelnde System im weiteren Projektverlauf evaluiert werden kann, da mindestens die intellektuell erschlossenen Relationen auch vom System erkannt werden sollten.

Die Datenbestände der Testkollektionen selbst weisen keine strukturellen Besonderheiten oder signifikante Korrelationen auf. Eine Vielzahl der Dokumente weist nur eine sehr geringe Dateigröße von weniger als 100 KB auf und lediglich einige Ausreißer heben die durchschnittliche Dateigröße stark an.

Versucht man allfällige Relationen zu bestimmen, so müssen zunächst Relationen der Dateien zueinander (z. B. Dateiformat, Erstellungsdatum) und inhaltlichen Relationen (z. B. gleiches Themengebiet) unterschieden werden. Nur bei inhaltlichen Relationen ist die Unterscheidung zwischen strukturierten und unstrukturierten Daten sinnvoll, da maschinelle Auswertbarkeit und somit auch die Qualität der Relationsextraktion bei strukturierten Daten zu wesentlich besseren Ergebnissen führt.

Dateispezifische Attribute können jederzeit automatisiert ausgewertet werden, führen aber nicht zwingend zu zuverlässigen Ergebnissen, da sie durch zahlreiche Prozesse wie das Brennen von Daten auf CDs (Attribut „Date Created" e. g.) beeinflusst werden können. Der Dateityp sowie das wechselseitige Beinhalten von Dateien (E-Mail-Attachments, ZIP-Dateien, etc.) können als weitestgehend zuverlässig und nachweisbar erachtet werden.

Inhaltliche Attribute sind schwerer zu bestimmen. Das Einbeziehen von Metadaten in die Auswertung der Dateien hat sich als wenig zuverlässig erwiesen, da diese oft nur rudimentär vorhanden oder auch fehlerbehaftet sind. Dementsprechend müssen inhaltliche Verknüpfungen der Dateien primär entweder aus der Dateistruktur oder auch aus den vorkommenden Termini gewonnen werden. Zeitliche und räum-

[1] *http://www.r-project.org/*, Stand 20.01.2007.

liche Bezüge können ebenfalls nur in einem geringen Teil der ausgewerteten Dateien eindeutig nachvollzogen werden.

## 5 Potenziell geeignete betriebliche Anwendungen zur Visualisierung geografisch referenzierter Daten

Die Analyse der Testkollektionen zeigt, zumindest exemplarisch, dass eine Verknüpfung von Sachdaten aus strukturierten bzw. unstrukturierten Unternehmensdatenbeständen mit räumlichen Daten aus Kartenmaterial nicht generell automatisierbar ist bzw. dass die von G10 entwickelte Technologie verfeinert und erweitert werden muss, um angemessene und verwertbare Ergebnisse zu liefern.

Die Auswertung geografischer Beziehungen ist insbesondere für Branchen wie z. B. Transport und Logistik, Reiseveranstaltern oder der Immobilienbranche oder im Rahmen der Fahndung von besonderem Interesse. Um auftretenden Problemstellungen entgegenzuwirken, werden inzwischen vermehrt BI-Software oder auch reine GI-Systeme eingesetzt, die zur Unterstützung des jeweiligen Kerngeschäftes dienlich sind. Diese Systeme benötigen allesamt jedoch stark strukturierte Daten aus Datenbank oder in spezifischen Formaten, um Zusammenhänge hervorheben zu können. Im Bereich der Suchmaschinen und Desktopsuchen dagegen gibt es ad hoc noch keinerlei Produkte, die ähnliche Ziele und Methoden anvisieren.

## 6 Konzept eines Visualisierungssystems

Die Wahl geeigneter Visualisierungstechniken hängt von der zu Grunde liegenden Datenstruktur ab. Als Ergebnis auf eine Suchanfrage ist zunächst eine netzartige Datenstruktur zu erwarten. Greift man jedoch einen Einzeltreffer heraus und sucht von diesem ausgehend weitere Treffer anhand direkter Relationen, so erhält man eine hierarchische Datenstruktur. Als gängige Visualisierungen für Netze sind zunächst Graphen anzuführen, wobei diese für die Visualisierung umfassender Datenmengen nur eingeschränkt geeignet sind. Hierarchien dagegen werden meist als Baumstruktur visualisiert, beispielsweise in Form eines Hyperbolic Trees [Lamping 1995], eines Cone Trees [Robertson 1991] oder einer Treemap [Johnson & Shneiderman 1991, Shneiderman 1992].

Um diese Visualisierungstechniken geeignet in die Benutzeroberfläche einzubetten, bedarf es einer geschickten Kombination an Interaktionstechniken. Sogenannte „Focus+Context"- oder auch Distortion-Techniques [Leung 1994] wie Fisheye

Views [Furnas 1986] und Bifocal Displays [Spence 1993] e. g. werden oftmals eingesetzt um „Overview and Detail" [Shneiderman 1996] auf einen Blick zu liefern. Weit verbreitet ist auch der Einsatz von Linking und Brushing, bei dem einzelne Elemente der Visualisierungen miteinander verknüpft interagieren und optisch hierbei hervorgehoben werden. Spätestens seit Raskins Vision einer „Zoomworld" (vgl. [Raskin 00]) werden auch Zooming-Techniken immer verstärkter angewandt.

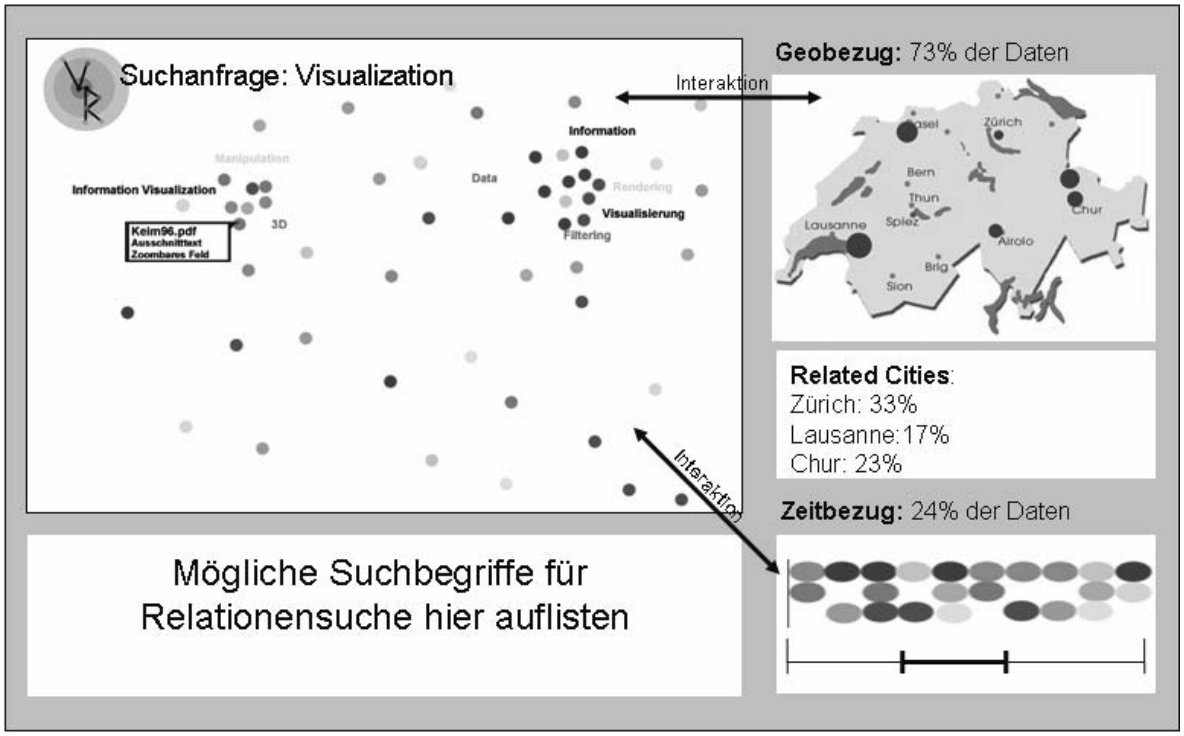


*Abbildung 1: Erster Entwurf der Benutzeroberfläche für eine Suchanfrage*

Abbildung 1 stellt eine Benutzeroberfläche zur Repräsentation von Suchergebnissen dar. Das Gerüst besteht zunächst aus drei elementaren Modulen: der Document Map (links oben) zur Visualisierung der Treffermenge, dem integrierten GI-System zur Visualisierung geografisch referenzierter Daten (rechts oben) und der Zeitleiste (rechts unten), die unter Verwendung der Fish-Eye-Technik umgesetzt werden soll. Weitere Felder können beispielsweise mit textuellen Ergänzungen oder zusätzlichen Visualisierungen genutzt werden.

Im dargestellten Beispiel ist der jeweilige Dateityp auf die visuelle Variable Farbe gemappt, während die Sättigung die Anzahl der Treffer darstellt. Zudem sind relevante Termini – wie bei den meisten Document Maps üblich – als Landmarken eingeblendet.

Im GI-System repräsentiert die Größe der Kreise die Anzahl der Treffer an einem Ort. Die einzelnen Bereiche interagieren per Linking (Verknüpfung gleicher Elemente in unterschiedlichen Darstellungen) und Brushing (optisches Hervorheben verknüpfter Elemente) miteinander.

Abbildung 2 zeigt die Benutzeroberfläche, die dem Nutzer bei der Suche nach Relationen dienlich sein soll, wobei das Grundgerüst äquivalent zum dargestellten Entwurf in Abbildung 1 aufgebaut wurde. Das Design orientiert sich grundsätzlich an der Metapher einer Zielscheibe, wobei die eingeblendeten Ringe lediglich der Orientierung dienen sollen und in ihrer Anzahl nicht den Grad der Relationen widerspiegeln. Per Mouse-Over-Effekt können Verbindungen zwischen den einzelnen Dateien, die wiederum farblich kodiert wurden, als Graph eingeblendet werden. Da erwartet wird, dass sehr große Datenmengen visualisiert werden müssen, wurde der Einsatz von Fish-Eye-Techniken im Bereich der Relationenvisualisierung als sinnvoll erachtet. Auch hier unterstützen die eingeblendeten Ringe den Nutzer darin, die jeweilige Verzerrung zu erkennen. Ab einem gewissen Grad an Relationen sollen diese nicht mehr direkt präsentiert werden. Pfeile entsprechender Größe oder geeignete Landmarken dienen dem Nutzer an dieser Stelle als Wegweiser, in welche Richtung er sich mittels Panning, dem Verschieben des betrachteten Ausschnitts der Visualisierung, weiterbewegen kann und welche Treffermenge in dieser Richtung zu erwarten ist.

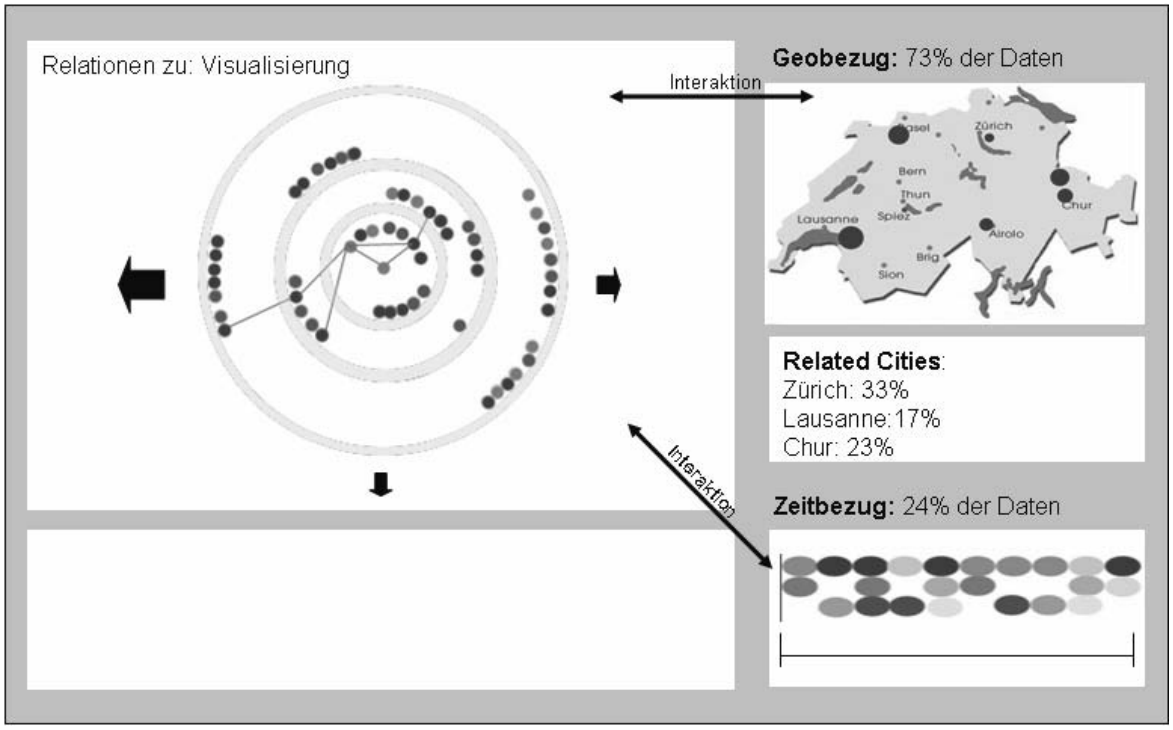


*Abbildung 2: Erster Entwurf der Benutzeroberfläche für die Relationensuche*

## 7 Ausblick

Unternehmensdatenbestände beinhalten, so das Ergebnis der durchgeführten Analyse, nur zum geringen Teil Dateien, die einen eindeutigen geografischen oder zeitli-

chen Bezug aufweisen. Insbesondere aus unstrukturierten Daten sind diese Attribute schwer zu gewinnen. Entsprechend werden im Rahmen des Projektes Visual Relations verschiedene Visualisierungen miteinander kombiniert, die dem Nutzer unterschiedliche Perspektiven auf die Treffermenge erlauben. Ein weiterer Vorteil des modularen Systemaufbaus liegt beispielsweise darin, Daten mit geografischen Bezug und Daten ohne eindeutigen geografischen Bezug gleichzeitig darstellen zu können und die jeweiligen Visualisierungen interaktiv miteinander zu verbinden. Es verbleibt dem Nutzer selbst im Anschluss an die Recherche, durch „Filtering"-Mechanismen einerseits oder durch „Zooming and Panning" andererseits, den dargestellten Informationsraum über den Informationszugriff seiner Wahl zu erkunden. Die eingesetzten Visualisierungstechniken müssen ebenso wie die Mechanismen zur Extraktion zeitlicher und örtlicher Relationen – nach einer Evaluation des Prototypen – noch genauer verfeinert werden. Insbesondere durch gezieltes Dato Preprocessing auf Basis einer Auswertung der vom System gefunden Relationen könnte die Qualität der integrierten Visualisierungen wesentlich verbessert werden. Dieser Beitrag führt zu folgenden Erkenntnissen: Geografische Metaphern haben ein Potential, um in der Unternehmenspraxis die Informationsanalyseprozesse zu optimieren, allerdings für eng fokussierte Anwendungsbereiche. Die Analyse der Unternehmensdatenbasis lässt durchaus Attribute extrahieren, die als Grundlage für neue Visualisierungsansätze dienen können. Die vorgestellten Prototypen versprechen einen Lösungsansatz zur skizzierten Problematik und eine verbesserte Interaktion der Nutzer mit Information Retrieval-Systemen zu liefern.

## 8 Literatur

[Balaouras et al. 2006], Balaouras Stephanie, Schreck Galen, Batiancila Rachel, Disk-Based Data Protection Forecast: 2006 To 2011, Enterprises Shift To Disk As The First Line Of Protection, Forrester Research, *http://www.forrester.com/Research/PDF/0,5110,40036,00.pdf, 17.11.2006*

[Dragoon 2003], Dragoon Alice, Business Intelligence Gets Smart(er), Sep. 15, 2003 Issue of CIO Magazine, 2003, http://www.cio.com/archive/091503/smart.html, Stand 26.10.06

[EMC 2006], Homepage der Firma EMC Switzerland, URL: *http://switzerland.emc.com/ilm/, Stand 12.12.2006*

[Feldman 2004], Feldman Susan: The high cost of not finding information, March 2004 Issue of KMWorld Magazine, 2004, *http://www.kmworld.com/Articles/ReadArticle.aspx?ArticleID=9534*, Stand 26.10.06

[Furnas 1986], Furnas, G. W. 1986. Generalized fisheye views. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Boston, Massachusetts,

United States, April 13 – 17, 1986). M. Mantei and P. Orbeton, Eds. CHI '86. ACM Press, New York, NY, 16-23.

[Johnson/ Shneiderman 1991], Johnson, B. and Shneiderman, B. 1991. Tree-Maps: a space-filling approach to the visualization of hierarchical information structures. In Proceedings of the 2nd Conference on Visualization '91 (San Diego, California, October 22 – 25, 1991). G. M. Nielson and L. Rosenblum, Eds. IEEE Visualization. IEEE Computer Society Press, Los Alamitos, CA, 284-291.

[Keim 2002], Keim, D. A. 2002. Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics 8, 1 (Jan. 2002), 1-8.

[Lamping 1995], Lamping, J., Rao, R., and Pirolli, P. 1995. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Denver, Colorado, United States, May 07 – 11, 1995). I. R. Katz, R. Mack, L. Marks, M. B. Rosson, and J. Nielsen, Eds. Conference on Human Factors in Computing Systems. ACM Press/Addison-Wesley Publishing Co., New York, NY, 401-408.

[Leung 1994], Leung, Y. K. and Apperley, M. D. 1994. A review and taxonomy of distortion-oriented presentation techniques. ACM Trans. Comput.-Hum. Interact. 1, 2 (Jun. 1994), 126-160

[Lyman/ Varian 2006], Lyman Peter, Varian Hal, How much information (2006), Studie über die weltweite Datenproduktion, *http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/,* Stand 12.12.06

[Raskin 2000], Raskin, J. 2000 The Humane Interface: New Directions for Designing Interactive Systems. ACM Press/Addison-Wesley Publishing Co.

[Robertson 1991], Robertson, G. G., Mackinlay, J. D., and Card, S. K. 1991. Cone Trees: animated 3D visualizations of hierarchical information. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology (New Orleans, Louisiana, United States, April 27 – May 02, 1991). S. P. Robertson, G. M. Olson, and J. S. Olson, Eds. CHI '91. ACM Press, New York, NY, 189-194.

[Schumann/Müller 2000], Schumann H., Müller W., (2000)Visualisierung: Grundlagen und allgemeine Methoden, Berlin

[Schneiderman 1992], Shneiderman, B. 1992. Tree visualization with tree-maps: 2-d space-filling approach. ACM Trans. Graph. 11, 1 (Jan. 1992), 92-99.

[Schneiderman 1996], Shneiderman, B. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In Proceedings of the 1996 IEEE Symposium on Visual Languages (September 03 – 06, 1996). VL. IEEE Computer Society, Washington, DC, 336.

[Spence 1993], Spence, R. 1993. A taxonomy of graphical presentation. In INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems (Amsterdam, The Netherlands, April 24 – 29, 1993). S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, and T. White, Eds. CHI '93. ACM Press, New York, NY, 113-114