Využití techniky náhodného indexování v oblasti detekce plagiátů

Zdeněk Češka

Katedra informatiky a výpočetní techniky, Fakulta aplikovaných věd
Západočeská univerzita v Plzni
Univerzitní 22, 306 14 Plzeň, Česká republika
zceska@kiv.zcu.cz

Abstrakt. *Rostoucí snaha plagiovat cizí práce, především v oblasti školství, zapříčinila vývoj nových a lepších metod, které by těmto intrikám čelily. Tento článek rozvíjí myšlenku aplikace Latentní sémantické analýzy (LSA) v oblasti detekce plagiátů a navrhuje nová vylepšení. Hlavním diskutovaným předmětem je aplikace kompresní techniky tzv. náhodného indexování, která transformuje data do alternativního zmenšeného prostoru. Kromě toho se článek zabývá normalizací podobností mezi dokumenty a přináší novou asymetrickou normalizační formuli. Experimenty byly provedeny na manuálně vytvořeném korpusu českých plagiátů, který obsahuje 1500 dokumentů o politice. Dosažené výsledky indikují, že kompresní technika dokáže významně snížit časové požadavky pro LSA. Aplikací nové normalizační formule lze navíc dosáhnout i vyšší přesnosti detekce plagiátů při současně nižších časových požadavcích.*

1 Úvod

Zvyšující se zájem o určování autorství psaných dokumentů vede k vývoji nových pokročilých metod, které jsou schopny automaticky detekovat případy plagiátorství. Tento problém je navíc umocněn množstvím volně dostupných dokumentů na Internetu, pojednávajících o různorodých tématech. Cílem metod pro detekci plagiátů je objektivně posoudit rozličné zdroje a nalézt ty, které byly nějakým způsobem plagiovány. Ačkoli současné moderní metody dávají dobré výsledky, stále je vyžadováno konečné lidské rozhodnutí o tom, co lze považovat za plagiát. Současné metody slouží především jako vodítko a významným způsobem šetří lidský čas.

Tento článek je zaměřen na zlepšení výsledků metody SVDPLAG [2], která je založena na Latentní sémantické analýze (LSA), viz [8]. Pro extrakci latentní sémantiky z textu se využívá matematická metoda Singulární hodnotové dekompozice (SVD - [1]). Klíčové příznaky, které jsou touto metodou zkoumány, představují fráze obsažené v textových dokumentech. Jak popisuje článek [2], fráze jsou reprezentovány slovními N-gramy, které se postupně analyzují a extrahují z předzpracovaného textu.

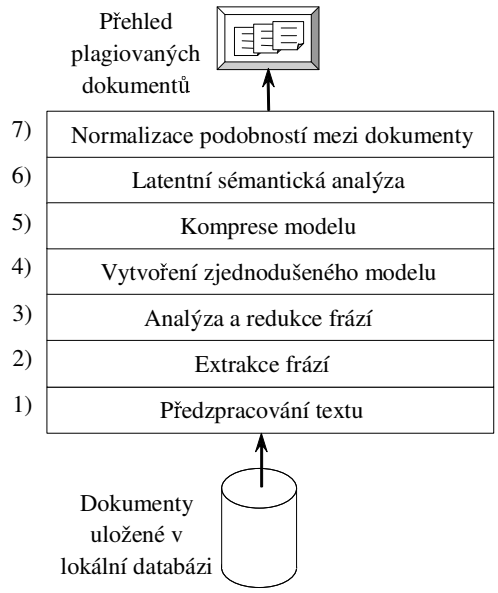
SVDPLAG má odlišný přístup k analýze textu oproti jiným metodám, pracujícím pouze s kosinovou mírou vektoru, která obsahuje počty výskytů jednotlivých slov, viz [12], podobně systém SCAM [13]. Rovněž metody založené na prostém průniku společných slovních N-gramů, jako je systém FERRET [9], nedosahují vyžadovaných výsledků. Bližší popis různých přístupů a jejich souhrn lze nalézt např. v článcích [3] a [10]. SVDPLAG jde v tomto ohledu cestou rozsáhlých statistických výpočtů v rámci LSA, jež provádí analýzu všech dokumentů současně. Podávané výsledky jsou proto podstatně vyšší než u ostatních metod. Nevýhodou jsou vyšší časové požadavky, které jsou hlavním předmětem řešení tohoto článku.

Další text v tomto článku je organizován následovně. Sekce 2 navrhuje užití kompresní techniky založené na náhodném indexování, společně s novou optimalizovanou normalizační formulí pro výpočet podobností mezi dokumenty. Sekce 3 prezentuje výsledky navržených modifikací na metodě SVDPLAG. Závěrečná diskuse dosažených výsledků je podána v sekci 4.

2 Navrhovaná vylepšení SVDPLAG metody

Princip této metody [2] je založen na LSA, kde se jako jádro pro extrakci sémantických vztahů využívá matematická metoda SVD. Na základě toho je též odvozen název metody pro detekci plagiátů SVDPLAG.

Obrázek 1 prezentuje jednotlivé vrstvy zpracování, které byly detailně popsány v článku [2]. Jediná významná modifikace spočívá v přidání vrstvy pro kompresi modelu, podstatně urychlující výpočet následujícího LSA. Rovněž byla provedena drobná úprava v sedmé vrstvě, zahrnující novou asymetrickou normalizační formuli. Provedené modifikace jsou popsány v následujícím textu.



Obrázek 1. Vrstvy zpracování SVDPLAG metody.

2.1 Komprese modelu metodou náhodného indexování

Navržený model fráze x dokument v článku [2] se při zpracování rozsáhlých kolekcí potýká s velkými rozměry

matice A. Nechť A je n x m obdélníková matice složená z n vektorů $[A_1, A_2, \ldots, A_n]$, kde vektor A_i představuje fráze obsažené v dokumentu i. Vektor A_i se skládá z m prvků $a_{i,j}$, kde každý prvek je váženou frekvencí výskytu fráze j v dokumentu i. Možné řešení problému rozměrné matice A skýtá kompresní technika, která transformuje A do alternativního prostoru obsahující přibližně stejnou informaci, s využitím menšího počtu dimenzí.

Kanerva a kol. navrhli techniku *náhodného indexování* [7], která využívá rozložení prvků v řídké matici [6]. Tato technika umožňuje podstatně zmenšit jeden z rozměrů matice A. Kanerva a kol. aplikovali tento postup na matici kódující vztahy slovo-dokument. Zmenšená matice byla následně použita pro výpočet podobnosti slov s pomocí LSA. S ohledem na detekci plagiátů je matice A, představující model výskytu frází ve zkoumaných dokumentech, extrémně řídká. V tomto ohledu lze techniku náhodného indexování dobře uplatnit.

Transformaci původní matice A o rozměru m x n do nového komprimovaného prostoru $A´$ s rozměrem $m´$ x n naznačuje rovnice (1).

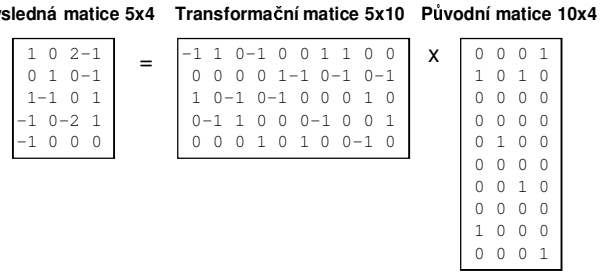
$$A' = T \times A \tag{1}$$

T je v tomto případě transformační matice $m´$ x m složená z m indexových vektorů $[T_1, T_2, \ldots, T_{m'}]$, kde každý indexový vektor T_i obsahuje o náhodně umístěných 1 a -1. Kromě toho jsou všechny indexové vektory vzájemně lineárně nezávislé. Počet náhodně umístěných 1 a -1 musí splňovat podmínku $o << m'$. Aplikací všech těchto kriterií získáme P-unitární matici T, která je pravděpodobnostní a splňuje podmínku (2).

$$\frac{T^T \times T}{2 \cdot o} \approx I \tag{2}$$

Výsledná matice $A´$ tudíž obsahuje dobrou aproximaci informace obsažené v původní matici A.

Obrázek 2 prezentuje náhodné indexování na příkladu. Transformační matice T byla vytvořena dle stanovených kritérií, kde o = 1. Původní matice A je velmi řídká, odpovídající situaci výskytu frází v dokumentech, kde řádky představují fráze a sloupečky jsou dokumenty. Transformací získáme matici $A´$, jejíž počet řádek byl podstatně zredukován v porovnání s A. Fráze jsou nyní namapovány na indexové vektory. Obsažená informace v novém prostoru je nicméně stále přibližně stejná, přinejmenším pro náš účel detekce plagiátů.



Obrázek 2. Příklad náhodného indexování.

2.2 Normalizace podobností mezi dokumenty

Normalizací podobností mezi dokumenty je nutné se zabývat z důvodu redukčního procesu ve třetí vrstvě, který odstraňuje méně významné fráze. Více informací o tomto problému lze nalézt v článku [2], kde byla rovněž uvedena formule pro tzv. symetrickou (SYM) normalizaci, viz (3). V této formuli $sim_{SVD}(R, S)$ představuje podobnost mezi dokumenty R a S vypočtenou na základě SVD procesu, $ph_{orig}(D)$ označuje množinu frází obsažených v dokumentu D před redukcí a $ph_{red}(D)$ je množina frází po redukci.

$$sim_{SYM}(R,S) = sim_{SVD}(R,S) \cdot \sqrt{\frac{|ph_{red}(R)|}{|ph_{orig}(R)|} \cdot \frac{|ph_{red}(S)|}{|ph_{orig}(S)|}} \tag{3}$$

Nevýhodou symetrické normalizace je nerelevantní hodnocení párů dokumentů s velmi rozdílnými velikostmi, kde jeden je podmnožinou druhého. Formule (4) řeší tento problém výběrem menší z dvou množin frází před redukcí. Tuto normalizaci proto nazýváme jako asymetrickou (ASYM). K odlišení těchto dvou modifikací, označujeme SVDPLAG jako $\text{SVDPLAG}_{\text{SYM}}$ nebo $\text{SVDPLAG}_{\text{ASYM}}$.

$$sim_{ASYM}(R,S) = sim_{SVD}(R,S) \cdot \frac{\sqrt{|ph_{red}(R)| \cdot |ph_{red}(S)|}}{\min\left(|ph_{orig}(R)|, |ph_{orig}(S)|\right)} \tag{4}$$

Dále zavádíme práh $\tau \in \langle 0,1 \rangle$, který představuje minimální stupeň plagiátorství. Pokud je výsledná podobnost mezi dokumenty R a S větší než τ, jsou oba z dokumentů považovány za plagiované, viz (5). Podobnostní míra sim může být v tomto případě zastoupena jak symetrickou, tak asymetrickou variantou.

$$plagiarized(R,S) = \begin{cases} true & if \quad sim(R,S) \geq \tau \\ false & if \quad sim(R,S) < \tau \end{cases} \tag{5}$$

3 Experimenty

3.1 Testovací data

Veškeré experimenty v tomto článku byly provedeny na korpusu 1500 plagiovaných dokumentů o politice psaných v českém jazyce. Celkově se tento korpus skládá z 550 dokumentů, které byly manuálně plagiovány studenty. K tomuto účelu bylo z ČTK korpusu [4], ročník 1999, vybráno 350 zpráv o politice, použitých jako podklad pro vytvoření plagiátů. Zbylých 600 dokumentů bylo vybráno ze stejného zdroje jako nezávislé zpráv na stejné téma, sloužící jako kontrola.

Pro vytvoření 550 plagiovaných dokumentů z 350 zdrojových dokumentů, byli studenti pověřeni kombinovat dva a více náhodně vybraných dokumentů. Výsledkem je, že každý dokument má odlišný stupeň podobnosti se zdrojovým dokumentem.

Při vytváření plagiátů byly uvažovány následující pravidla:

1. Zkopíruj několik odstavců z vybraných dokumentů
2. Smaž okolo 20% vět z nově vytvořeného dokumentu
3. Smaž okolo 10% slov s uvážením smysluplnosti vět
4. Zaměň okolo 20% vět z různých odstavců

5. Přeformuluj okolo 10% vět, přidáním nových myšlenek do textu
6. Pro zajištění smysluplnosti textu mohou být vložena nebo modifikována některá slova v textu

Pro speciální účely byl vytvořen zmenšený korpus 500 dokumentů, který je pouze zmenšenou verzí původního korpusu. Celkově tento zmenšený korpus obsahuje 170 plagiovaných dokumentů, 173 zdrojových dokumentů z ČTK a 157 nezávislých zpráv o politice jakožto kontrola.

3.2 Testovací prostředí

Veškeré experimenty byly provedeny na Intel Core 2 Duo E6600, 4 GB RAM a operačním systému Windows Server 2003 R2 v 64-bitovém režimu. Naše experimentální prostředí bylo vyvinuto v .NET Framework 3.5 s využitím Extreme Optimization Numerical Libraries v3.1 [5]. Pro efektivní měření časových požadavků byl umožněn běh pouze jednoho vlákna.

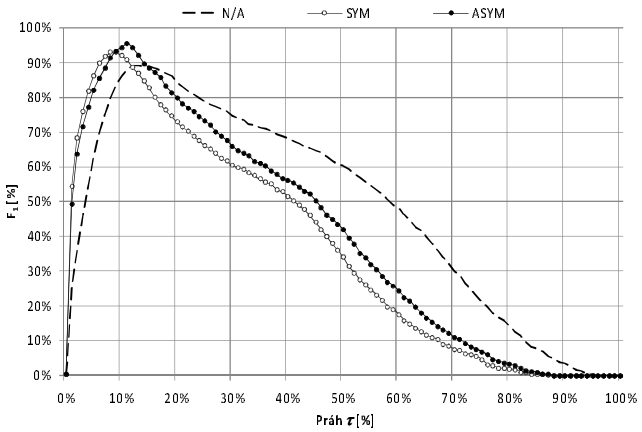
3.3 Užité metriky

K porovnání naměřených výsledků zavádíme standardní míru přesnosti p a úplnosti r dle [11]. Dále zavádíme míru F_1, která kombinuje přesnost a úplnost do harmonické střední hodnoty

$$F_1 = \frac{2 \cdot p \cdot r}{p + r}. \qquad (6)$$

3.4 Vliv normalizace podobností mezi dokumenty

Obrázek 3 prezentuje rozdíl mezi symetrickou (SYM) a asymetrickou (ASYM) normalizací, kde je zachycena závislost míry F_1 na prahu τ. Asymetrická normalizace získává významnou výhodu pro dokumenty nestejné délky, kde jeden je podmnožinou druhého, což se odráží na výsledku 95,68% F_1 oproti symetrické normalizaci 93,43% F_1.



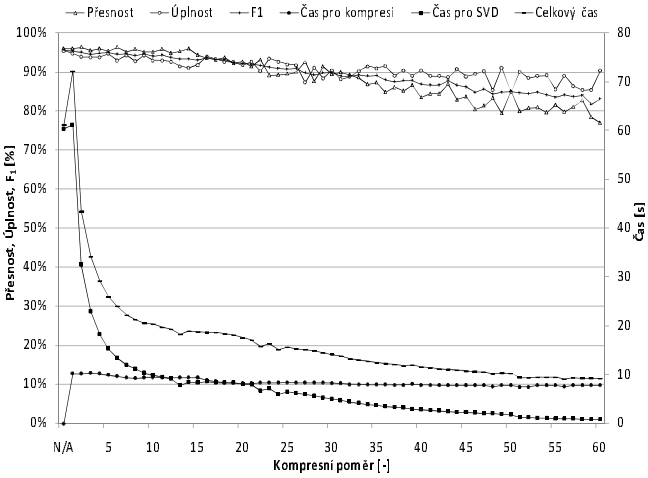
Obrázek 3. Závislost míry F_1 na prahu τ pro metodu SVDPLAG (obě varianty symetrická i asymetrická normalizace) s využitím 4-gramů jako příznaků.

Na obrázku je rovněž zachycena křivka bez aplikace normalizační formule „N/A“. V tomto případě dosahuje F_1 pouhých 89,34%, což naznačuje důležitost normalizačního procesu, je-li ve třetí vrstvě aktivován redukční proces odstraňující méně významné fráze. Z důvodu významné výhody asymetrické normalizace jsou následující testy provedeny pouze na této variantě.

3.5 Vliv náhodného indexování

Technika komprese příznaků je klíčovou součástí SVDPLAG metody, která podstatným způsobem snižuje časové i paměťové požadavky pro SVD a umožňuje zpracování rozsáhlých dat.

Obrázek 4 zachycuje chování pro různé kompresní poměry na plném korpusu 1500 dokumentů. Přesnost a úplnost naznačují maximální výchylky způsobené náhodným indexováním. Ze statistického pohledu je efekt takový, že pokles v přesnosti vyvolá nárůst v úplnosti. Vlastní míra F_1, která je střední harmonickou hodnotou mezi přesností a úplností, pak pouze zvolna klesá s rostoucími kompresní poměry.



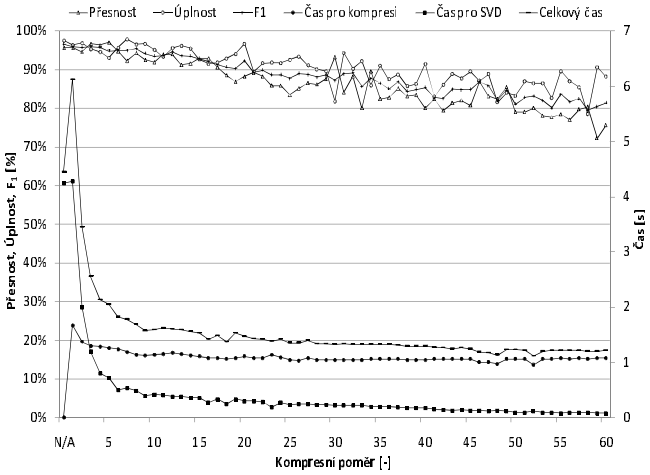
Obrázek 4. Vliv kompresního poměru na míru F_1 a výpočetní čas metody SVDPLAG$_{ASYM}$. Experiment byl proveden na plném korpusu 1500 dokumentů.

Situace, kdy není komprese aplikována, je v obrázku označena „N/A“. V tomto případě SVD vyžaduje k běhu 60,34 vteřin. Aktivace vlastní komprese vyžaduje dodatečných 10,20 vteřin pro překódování matice, nicméně vyšší kompresní poměry významně snižují čas pro SVD. Kupříkladu kompresní poměr 1:10 snižuje výpočetní čas SVD na 10,02 vteřin, při současném poklesu času pro vlastní kompresi na 9,48 vteřin. Výsledný čas pro transformaci matice do alternativního prostoru společně s SVD procesem je 19,50 vteřin, což je třikrát nižší čas v porovnání s původními 60,34 vteřinami. Po uplatnění kompresního poměru 1:10 klesá F_1 míra na 94,70% z původních 95,68% (bez komprese pro variantu s asymetrickou normalizací). Vyšší kompresní poměry přináší další snížení časových požadavků, avšak rovněž větší výchylky v přesnosti a úplnosti, vedoucí k výraznějšímu poklesu F_1 míry.

Obrázek 5 prezentuje chování na zmenšeném korpusu 500 dokumentů. Technika náhodného indexování dosahuje statisticky lepších výsledků pro rozměrnější data, což můžeme odvodit porovnáním s předchozím obrázkem pro plný korpus 1500 dokumentů. Pro menší objemy dat lze očekávat větší výchylky v přesnosti i úplnosti a rovněž

prudší pokles F_1. Naopak u rozsáhlejších dat je evidentní podstatně hladší průběh, který dovoluje užití vyšších kompresních poměrů.

V obou experimentech byl počet náhodně umístěných 1 a -1 stanoven na 10, viz Sekce 2.1. Tímto je splněna podmínka, že počet náhodně umístěných prvků musí být podstatně menší než rozměr dimenze m' po kompresi. Během experimentů jsme vyzkoušeli široký počet náhodně umístěných čísel, avšak nepodařilo se nám odhalit žádný významný vliv na přesnost ani úplnost.



Obrázek 5. Vliv kompresního poměru na míru F_1 a výpočetní čas metody SVDPLAG$_{ASYM}$. Experiment byl proveden na zmenšeném korpusu 500 dokumentů.

4 Závěr

Tento článek představil techniku náhodného indexování a její aplikaci v oblasti detekce plagiátů psaného textu, konkrétně na metodě SVDPLAG. Tato metoda je založena na Latentní sémantické analýze (LSA), využívající matematickou metodu Singulární hodnotové dekompozice (SVD) pro extrakci latentní sémantiky z analyzovaného textu. Extremní řídkost zkoumané matice, kódující vztahy fráze-dokument, dovoluje její kompresi technikou náhodného indexování. Tato technika transformuje původní matici do alternativního prostoru. Tímto postupem lze významně urychlit zpracování rozsáhlých datových kolekcí.

Rovněž byla navržena nová asymetrická normalizace podobností mezi dokumenty, která výrazným způsobem zlepšuje ohodnocení dokumentů nestejné délky, kde jeden je podmnožinou druhého.

Tabulka 1. Přehled dosažených výsledků pro SVDPLAG.

Normalizace	Nastavení	Práh τ [%]	F_1 [%]	Celkový čas [s]
SYM	k.p. = n/a	9,3	93,43	60,34
ASYM	k.p. = n/a	11,0	95,68	60,34
ASYM	k.p. = 10	10,8	94,70	19,50

Tabulka 1 shrnuje dosažené výsledky pro vlastní kompresní techniku i normalizaci. Jak můžeme vidět, symetrická (SYM) normalizace dosahuje pouhých 93,43% F_1, kdežto asymetrická (ASYM) 95,68% F_1. Aplikací techniky náhodného indexování s kompresním poměrem k.p. = 10 klesá F_1 na 94,70%, nicméně současně se podstatně zrychluje SVD výpočet. V našem případě klesají časové požadavky pro kompresi a SVD proces na jednu třetinu, z 60,34 vteřin na 19,50.

Navržené úpravy vylepšují původní SVDPLAG metodu jak po stránce vyšší F_1 míry, tak po stránce nižších časových požadavků. Dalšího snížení časových požadavků by bylo možné dosáhnout paralelním zpracováním.

Poděkování

Tato práce byla částečně podporována z prostředků Národního Programu Výzkumu II, projekt 2C06009 (COT-SEWing).

Reference

[1] M. Berry, S. Dumais, G. O'Brein, „Using Linear Algebra for Intelligent Information Retrieval", *SIAM Review*, vol. 37 issue 4, pp. 573-595, Society for Industrial and Applied Mathematics, Philadelphia, USA, 1995. ISSN 0036-1445.

[2] Z. Ceska, „Využití moderních přístupů pro detekci plagiátů", Proceedings of the ITAT 2008, Information Technologies – Applications and Theory, Hrebienok, Slovakia, pp. 23-26, 2008. ISBN 978-80-969184-8-5.

[3] P. Clough, „Plagiarism in natural and programming languages: An overview of current tools and technologies", *Internal Report CS-00-05*, Department of Computer Science, University of Sheffield, 2000.

[4] ČTK, „Czech News Agency: Political and General News Service". URL: http://www.ctk.cz/sluzby/databaze/dokumentacni/

[5] Extreme Optimization, „Numerical Libraries for .NET Professional 3.1 64 bit", poslední změna 24.11.2008. URL: http://www.extremeoptimization.com/

[6] P. Kanerva, „Sparse Distributed Memory", *The MIT Press*, 2nd rev. edition, 1988. ISBN 0-262-11132-2.

[7] P. Kanerva, J. Kristoferson, A. Holst, „Random Indexing of Text Samples for Latent Semantic Analysis", *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pp. 103-106, 2000.

[8] T. Landauer, P. Foltz, D. Laham, „An introduction to Latent Semantic Analysis", *Discourse Processes*, vol. 25, pp. 259-284, 1998.

[9] P. Lane, C. Lyon, J. Malcolm, „Demonstration of the ferret plagiarism detektor", *Proceedings of the 2nd International Plagiarism Conference*, Newcastle, 2006.

[10] H. Maurer, F. Kappe, B. Zaka, „Plagiarism – A survey", *Journal of Universal Computer Science*, vol. 12 issue 8, pp. 1050-1084, 2006.

[11] C. J. van Rijsbergen, „Information Retrieval", *Butterworth-Heinemann*, 2nd rev. edition, 1979. ISBN 0-408-70929-4.

[12] P. Runeson, M. Alexanderson, O. Nyholm, „Detection of duplicate defect reports using natural language processing", *Proceedings of the IEEE 29th International Conference on Software Engineering*, pp. 499-510, 2007.

[13] N. Shivakumar, H. Garcia-Molina, „SCAM: A copy detection mechanism for digital documents", *Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries*, Austin, 1995.