Âåêòîðíàÿ ìîäåëü ïðåäñòàâëåíèÿ òåêñòîâîé èíôîðìàöèè

Àâòîð Ñòàíèñëàâ Âàñèëüåâè÷ Ìî÷åíîâ
18.07.2008 ã.
Ïîñëåäíåå îáíîâëåíèå 20.07.2008 ã.

Òåêñò ïå÷àòíîãî èçäàíèÿ â ôîðìàòå PDF

Ââåäåíèå

Øèðîêîå ïðèìåíåíèå ñðåäñòâ âû÷èñëèòåëüíîé òåõíèêè â ðàçëè÷íûõ îáëàñòÿõ çíàíèé ñïîñîáñòâóåò áûñòðûì ðîñòîì îáúåìîâ îáðàá
èíôîðìàöèè.

Îñîáóþ àêòóàëüíîñòü ïðèîáðåòàåò ðàçðàáîòêà ìåòîäîâ èçâëå÷åíèÿ è ôîðìèðîâàíèÿ íîâûõ çíàíèé, íåîáõîäèìûõ äëÿ ðåøåíèÿ êîíêðåòí
êîìïëåêñíîå ñèñòåìíîå èñïîëüçîâàíèå ðàçëè÷íûõ ëèíãâèñòè÷åñêèõ ïîäõîäîâ è ìåòîäîâ èñêóññòâåííîãî èíòåëëåêòà, íàïðàâëåííûõ íà ì
èäåþ, çàëîæåííóþ àâòîðîì.

Îñíîâíîé çàäà÷åé, âîçíèêàþùåé ïðè ðàáîòå ñ ïîëíîòåêñòîâûìè áàçàìè äàííûõ, ÿâëÿåòñÿ çàäà÷à ïîèñêà äîêóìåíòîâ ïî èõ ñîäåðæàíèþ
ïîèñêîâûìè ìàøèíàìè â Èíòåðíåòå, çà÷àñòóþ íå îáåñïå÷èâàþò àäåêâàòíîãî âûáîðà èíôîðìàöèè ïî çàïðîñó ïîëüçîâàòåëÿ.

Ïîëüçîâàòåëü íå âñåãäà ìîæåò òî÷íî ñôîðìóëèðîâàòü ïîèñêîâûé çàïðîñ íà ïîëó÷åíèå èíôîðìàöèè, êîòîðàÿ åìó íåîáõîäèìà. Áîëåå òîãî,
Òðóäíîñòè, ñâÿçàííûå ñ ðåøåíèåì ýòîé çàäà÷è, çàêëþ÷àþòñÿ â ìíîãîîáðàçèè âîçìîæíûõ ôîðì âûðàæåíèÿ îäíîé è òîé æå èäåè, ìûñëè

Â äàííîé ñòàòüå ðàññìàòðèâàþòñÿ íåêîòîðûå ïîäõîäû ê ðåøåíèþ óêàçàííûõ ïðîáëåì. Îñíîâíîå âíèìàíèå óäåëåíî ìåòîäàì âåêòîðíîãî ïðå

1. Îáçîð ìåòîäîâ âåêòîðíîãî ïðåäñòàâëåíèÿ òåêñòà

Â êîíöå 80 õ ãîäîâ â ðàáîòàõ Ñàëòîíà [Salton et al. 1994] áûëà ïðåäëîæåíà âåêòîðíàÿ ìîäåëü êàê àëüòåðíàòèâà ëåêñè÷åñêîìó áå
ñïåêòðà ñëîâ è ñîîòâåòñòâóþùåãî âåêòîðà â ëåêñè÷åñêîì ïðîñòðàíñòâå. Â ïðîöåññå ïîèñêà ÷àñòîòíûé ïðîôèëü çàïðîñà ðàññìàòðèâà
ðåëåâàíòíûå äîêóìåíòû.

Â áîëåå ïðîäâèíóòûõ âåêòîðíûõ ìîäåëÿõ ðàçìåðíîñòü ïðîñòðàíñòâà ñîêðàùàåòñÿ ïóòåì ðàññìàòðèâàíèÿ íàèáîëåå ðàñïðîñòðàíåííûõ èëè ðå

Ãëàâíûì äîñòîèíñòâîì âåêòîðíîé ìîäåëè ÿâëÿåòñÿ âîçìîæíîñòü ïîèñêà è ðàíæèðîâàíèÿ äîêóìåíòîâ ïî ïîäîáèþ, òî åñòü ïî èõ áëèçîñòè
óäîâëåòâîðèòåëüíûìè, ÷òî îñíîâíîé ïðîáëåìîé, êîãäà çàïðîñ ñîäåðæèò ìàëî êëè÷åñòâî ñëîâ. Äëÿ ïîëó÷åíèÿ ëó÷øåé ðåëåâàíòíî
Latent Semantic Indexing (LSI). Ìîäåëü èñïîëüçîâàëà Singular Value Decomposition (SVD) äëÿ ïðåîáðàçîâàíèÿ ìàòðèöû

LSI ïîêàçàëà çíà÷èòåëüíîå ïðåâîñõîäñòâî â ðåçóëüòàòàõ ïîèñêà ïî ñðàâíåíèþ ñ ëåêñè÷åñêèì ìåòîäîì, îïèñàííûì ïîäåëè ÷àñ
òåõíèêîé [Salton 1989]. Îäíà èç íàèáîëåå ðàáîòîñïîñîáíûõ ñèſòåì íà îñíîâå LSI áûëà ſîçäàíà â Áåðêëè â 1995 ãîäó ìàéêëîì Áåððè

Îïèſûâàåìàÿ íèæå ſèſòåìà èſïîëüçóåò ſîâåðøåííî äðóãóþ èíòåðïðåòàöèþ ïîíÿòèÿ âåêòîðíîé ïîäëè òåſòà, â êîòîðîé íå ïðèìåíÿþò

2. Âåêòîðíàÿ ïîäëü ïðåäſòàâëåíèÿ òåſòîâîé èíôîðìàöèè

Â äàííîé ðàáîòå òåſò ðàſſìàòðèâàåòſÿ êàê ſòðóêòóðà, òî åſòü êàê ſîâîêóïſíòü îòäåëüíûõ âçàèſſâſçàííûõ ïðåäëîæåíèé, îá
àâòîſſ, ÷åðåç ſſſæåſòâî ſſäöåëåé ðàçſſãî ðàſſãà, ſſðåäſåëÿåſſûõ îòäåëüſſûſſ ſſðåäëîæåſèÿſſ, àáſçàöàſſ, ſſàðàãðàôàſſ, ãëàâàſſ è ò. ſ

Íèæå ſſðèâîäÿòſſ ſſðèſåðû ſſàòåſſàòè÷åſſêé èſſòåðſſðåòàöſſè âåêòſſðſſîé ſſîäëè íåſſòſſðſſãî çàſſſſ÷åſſſſãî ýëåſſåſſòà òåſſòà, ſſſſòſſÿùåſſ, íà

G = {G1, G2 , …, Gi, …, Gn}

Vg = {Vg1, Vg2 ,…. ,Vgi,, …, Vgn},

ãää G — ſſſæåſòâſſ ãëàâ; Gi — i àÿ ãëàâà, i = 1… n; Vg — ſſſæåſòâſſ âåêòſſðſſâ öåëåé ãëàâ; Vgi &mdas

Â ſſâſþ ſ÷åðåäü

Ai = {Ai1, Ai2 ,…, Aij ,…, Am}

Vài = {Vài1, Vài2, …, Vàij ,…, Vàim},

ãää Ài — ìíîæåñòâî àáçàöåâ i îé ãëàâû; Àij — j ûé àáçàö i îé ãëàâû, j = 1 … m; Vài — ìíîæåñòâî âåêòî
ãëàâû.

Ìàòåìàòè÷åñêàÿ èíòåðïðåòàöèÿ âåêòîðíîé ìîäåëè äëÿ ïðåäëîæåíèé âûðàæàåòñÿ â âèäå:

Pij = {Pij1, Pij2 ,…, Pijh ,…, Pk}

Vðij = {Vðij1, Vðij2, …, Vðijh,..., Vðijk},

ãää Ðij — ìíîæåñòâî ïðåäëîæåíèé i îé ãëàâû j ãî àáçàöà; Ðijh — h îå ïðåäëîæåíèå i îé ãëàâû j ãî àáçàöà, h = 1 &he
Vpijh — âåêòîð öåëè h îãî ïðåäëîæåíèÿ i îé ãëàâû j ãî àáçàöà.

Èç ïðåäñòàâëåííîãî îïèñàíèÿ âèäíî, ÷òî êàæäîìó ýëåìåíòó (ôðàãìåíòó) òåêñòà ñòàâèòñÿ â ñîîòâåòñòâèå íåêîòîðûé âåêòîð öåëè.

Êàê èçâåñòíî, ñìûñëîâûì è ãðàììàòè÷åñêèì öåíòðîì ïðåäëîæåíèÿ îáû÷íî ÿâëÿåòñÿ ñêàçóåìîå, âûðàæåííîå ãëàãîëîì (ïîëíîçíà÷íûì èëè
çàâèñèìûì ñëîâàìè èëè áåç íèõ) ìîãóò âûñòóïàòü â êà÷åñòâå çàêîí÷åííîãî ïðåäëîæåíèÿ. Ïðèìåðàìè òàêèõ ïðåäëîæåíèé, íàçûâàåì

Äâàäöàòü ïåðâîå. Íî÷ü. Ïîíåäåëüíèê. Î÷åðòàíüÿ ñòîëèöû âî ìãëå. (Àõìàòîâà).

Êðîìå èìåííûõ ïðåäëîæåíèé ìæíî ðàññìàòðèâàòü è íåïîëíûå, êîòîðûå îáðàçóþòñÿ èç ïîëíûõ ïóòåì ïîðåäåëåííûõ ñîêðàùåíèé. Íàïðèìåð

Îòâåðòêó! (âìåñòî Äàé îòâåðòêó!).

Ïðè÷èíû òàêîãî ñîêðàùåíèÿ, íàçûâàåìîãî ýëëèïñèñîì, ìîãóò áûòü ðàçíîîáðàçíû, íî îáû÷íî ñîêðàùàåòñÿ òà ÷àñòü ïðåäëîæåíèÿ, êîòîðà

Êàê óæå îòìå÷àëîñü âûøå, êàæäîå ïðåäëîæåíèå íåñåò â ñåáå ïîðåäëåííûé ñìûñë, çàêëàäûâàåìûé àâòîðîì, è îáåñïå÷èâàò ïðîâ
Â îáùåì ñëó÷àå êàæäîå ïðåäëîæåíèå èìååò ñîîòâåòñòâóþùèé âåêòîð öåëè. Òàêèì îáðàçîì, òåêñò ìîæíî ïîðàäëèòü êàê ñòðóêòó

Ïðåäëîæåíèÿ ìîäåëü âåêòîðà öåëè ìîæåò áûòü ïðåäñòàâëåíà â âèäå òðåõ êîìïîíåíò:

Vbegin — íà÷àëüíàÿ öåëü, âûðàæàåìàÿ ÷åðåç íà÷àëüíûé âåêòîð ñ çàäàííûìè êîíöåíòðàòàìè;

Vend êîíå÷íàÿ öåëü, âûðàæàåìàÿ ÷åðåç êîíå÷íûé âåêòîð ñ çàäàííûìè êîíöåíòðàòàìè;

Z âèä ñâÿçè ìåæäó íà÷àëüíûì âåêòîðîì è êîíå÷íûì âåêòîðîì .

Â êà÷åñòâå êîíöåíòðà ìîãóò âûñòóïàòü îòäåëüíûå ñëîâà, ïîíÿòèÿ, èìåííûå ãðóïïû, îòäåëüíûå ïðåäëîæåíèÿ, ñìûñëîâûå ãðó

Ïîñêîëüêó èìåþòñÿ òðè ñîñòàâëÿþùèå, êîòîðûå îïðåäåëÿþ âåêòîð, òî ñîîòâåòñòâåííî äëÿ ïîñëåäóþùåãî àíàëèçà íàìè áûëè âûäåë

1) ïðîñòîé âåêòîð: = () èëè V = ();

2) íóëåâîé âåêòîð: = (Ø);

3) ïîëíûé âåêòîð: = (,) ñî ñâÿçüþ Z;

4) ïóñòîé âåêòîð: = (,) áåç ñâÿçè Z;

5) ëåâûé âåêòîð: = ();

6) ïðàâûé âåêòîð: = ().

Â ñâîþ î÷åðåäü âåêòîðû è ìîãóò ñîñòîÿòü èç ïîäâåêòîðîâ, êàê îòäåëüíûõ ñàìîñòîÿòåëüíûõ ÷àñòåé êîðäèíàò, ïðèíàäëåæàùèõ ò

Êàæäàÿ êîîðäèíàòà èìååò ñâîè àòðèáóòû atr. Àòðèáóòàìè ìîãóò ÿâëÿòüñÿ âðåìåííûå èëè ïðîñòðàíñòâåííûå õàðàêòåðèñòèêè êîî

Ñîñòàâ êîîðäèíàò âåêòîðà îïðåäåëÿåòñÿ ñëîæíîñòüþ ïîñòðîåíèÿ ïðåäëîæåíèÿ. Â îáùåì ñëó÷àå ñîîîä÷èíåííîñòü îòäåëüíûõ ÷àñòåé ï

Ðàññìîòðèì ïðèìåíåíèå îïèñàííîé âûøå âåêòîðíîé ìîäåëè íà êîíêðåòíîì ïðèìåðå.

Ïðîàíàëèçèðóåì ñëåäóþùåå ïðåäëîæåíèå.

Âî âñå âðåìåíà ëþäè ñòàëêèâàþòñÿ ñ îäíèìè è òåìè æå ïðîáëåìàìè ýêîíîìèêè.

Äàííîå ïðåäëîæåíèå ìîæåò áûòü ïðåäñòàâëåíî â âåêòîðíîé ôîðìå: Vp(x1; y1) ñî ñâÿçüþ âèäà z1, èëè â óïðîùåííîé ôîðìå Vp(x1; y1) (

ãääå êîîðäèíàòà x1 = {ëþäè};

êîîðäèíàòà y1 = {ïðîáëåìû ýêîíîìèêè};

ñâÿçü âèäà z1 = (ñòàëêèâàþñÿ).

Ïðè ýòîì àòðèáóòàìè êîíöàêòà õ1 ÿâëÿþñÿ atrx = (âî âñå âðåìåíà), à àòðèáóòàìè êîíöàêòà y1 ÿâëÿþñÿ atry = (îäíè è òå æå)

Íà îñíîâå ââåòíðíîãî ïðåäñòàâëåíèÿ ìîäó áûëè ðåøåíû íåêîòîðûå ïðîáëåìû îáðàáîòêè òåêñòîâîé èíôîðìàöèè, â ÷àñòíîñòè:

– ñîêðàùåíèå îáúåìà èñõîäíîé èíôîðìàöèè äëÿ âûïîëíåíèÿ ïðîöåäóð àíàëèçà òåêñòà è ôîðìèðîâàíèÿ ñèñòåì è áàç çíàíèé;

– ñèíòåç òåêñòà ñ èñïîëüçîâàíèåì èíôîðìàöèè, èçâëåêàåìîé èç áàç çíàíèé.

Â ñëåäóþùåì ïàðàãðàôå ðàññìàòðèâàåòñÿ ãåíåòè÷åñêàÿ èíòåðïðåòàöèÿ òåõíîëîãèè ââåòíðíîãî ïðåäñòàâëåíèÿ òåêñòà.

3. Ïðèìåíåíèå òåõíîëîãèè ââåòíðíîãî ïðåäñòàâëåíèÿ ïðè àíàëèçå è ñèíòåçå òåêñòîâîé èíôîðìàöèè

Ðàññìîòðåííàÿ âûøå ââåòíðíàÿ ìîäåëü ïðåäñòàâëåíèÿ òåêñòîâîé èíôîðìàöèè ìîæåò áûòü èñïîëüçîâàíà ïðè àíàëèçå è ñèíòåçå òåêñ

Íà ðèñ. 1 ïðåäñòàâëåíà óïðîùåííàÿ èíòåðïðåòàöèÿ ââåòíðíîãî ïðåäñòàâëåíèÿ òåêñòà â ïðîñòðàíñòâå òðåõ êîîðäèíàò õ, y, atr.

Ðèñ. 1. Âåêòîðíîå ïðåäñòàâëåíèå òåêñòà

Íà ïðåäñòàâëåííî ðèñóíêå ïîêàçàíû òðè âåêòîðà Vp1(x1, y1) (z1); Vp2(x2, y2) (z2); Vp3(x3, y3) (z3) è èõ ïðîåêöèÿ íà ïëîñêîñòü x, ïëîñêîñòè: (x, y), (x, atr), (y, atr).

Àòðèáóòû ìîäóëè âðåìåííûå, ïðîñòðàíñòâåííûå è äðóãèå èçìåðÿåìûå õàðàêòåðèñòèêè.

Êîîðäèíàòû õi ïîðåäåëÿþò íà÷àëüíûå êîîðäèíàòû âåêòîðà. Êîîðäèíàòû yi ïîðåäåëÿþò êîíå÷íûå êîîðäèíàòû âåêòîðà.

Èñõîäÿ èç ïðåäûäóùåãî ïîèñàíèÿ âåêòîð V îïðåäåëÿåò êîíå÷íóþ öåëü ðàññìàòðèâàåìîé äèíèöû òåêñòà è èìåò ñòðóêòóðó âåêò

Ðàññìîòðèì èñïîëüçîâàíèå îïèñàííîé ìîäåëè íà ïðèìåðå.

Ïóñòü çàäàíû òðè âåêòîðà:

Vp1(x1, y1) (z1); Vp2(x1, y2) (z2); Vp3(x1, y3) (z3).

Íà ðèñ. 2 ïîêàçàíû íåêîòîðûå âîçìîæíûå âàðèàíòû ãåîìåòðè÷åñêîé èíòåðïðåòàöèè âçàèìîäåéñòâèÿ òðåõ âåêòîðîâ.

à

á

â

Ðèñ. 2. Ãåîìåòðè÷åñêàÿ èíòåðïðåòàöèÿ âçàèìîäåéñòâèÿ òðåõ âåêòîðîâ

Ïðèìåð, ïðåäñòàâëåííûé íà ðèñ. 2à, ïîêàçûâàåò, ÷òî êîîðäèíàòà õ1 âåêòîðà ïðåäñòàâëÿåò ñîáîé èåðàðõèþ ïîíÿòèé. Íàïðèìåð, õàðà

Íà ðèñóíêå 2á ïðåäñòàâëåíà ãåîìåòðè÷åñêàÿ èíòåðïðåòàöèÿ âçàèìîäåéñòâèÿ äðóãîé ãðóïïû èç òðåõ âåêòîðîâ:

Vp1(x1, y1) (z1); Vp2(x2, y2) (z2); Vp3(x3, y3) (z3),

ãäå y1 = x2, y2 = x3.

Ôàêòè÷åñêè ñîâîêóïíîñòü ýòèõ òðåõ âåêòîðîâ îïðåäåëÿåò íåêîòîðûé ðåçóëüòèðóþùèé âåêòîð V(x1, y3) (z’), ñîîòâåòñòâó

Ñëåäóþùèé ïðèìåð (ðèñ. 2â) èëëþñòðèðóåò íåçàâèñèìîñòü öåëåé â ïðèâåäåííîì íàáîðå âåêòîðîâ:

Vp1(x1, y1) (z1); Vp2(x1, y2) (z2); Vp3(x3, y3) (z3),

ãääå y2 = x3.

Äðóãèì ïðèìåíåíèì âåêòîðíîé ìîäåëè ÿâëÿåòñÿ âîçìîæíîñòü ðåàëèçàöèè ñèíòåçà òåêñòîâîé èíôîðìàöèè.

Ïðåäïîëîæèì, ÷òî ðåøàåòñÿ çàäà÷à, ñâÿçàííàÿ ñ ðàñêðûòèåì ïîíÿòèÿ õ1. Â ýòîì ñëó÷àå, âåêòîð öåëè äëÿ îïèñàíèÿ îïðåäåëåííûõ ïð
êîíñòóêöèîíàëüíûõ ïîäöåëåé.

Äîïóñòèì, â áàçå çíàíèé ïîíÿòèå õ1 îïðåäåëåíî íà ìíîæåñòâå îíòîëîãèé ÷åðåç âåêòîð Vp0(x1, y1) (z1). Â ñâîþ î÷åðåäü ïîâåêòîð y1
ðàñêðûòèÿ ïîíÿòèÿ õ1. Ìåõàíèçì ðàçâåðòûâàíèÿ âåêòîðà äëÿ îïèñàíèÿ ïðîöåññîâ è ÿâëåíèé ìîæåò áûòü äâîÿêèì: ëèáî íà îñíîâå ñîõ
ÿçûêå.

Íà îñíîâå ïðåäëîæåííîé ìîäåëè ðàçðàáîòàíà òåõíîëîãèÿ îáðàáîòêè òåêñòîâîé èíôîðìàöèè íà îñíîâå âåêòîðíîé ìîäåëè òåêñòà.

Çàêëþ÷åíèå

Ðàññìîòðåííûå â äàííîé ñòàòüå îñíîâíûå ïîëîæåíèÿ òåõíîëîãèè âåêòîðíîãî ïðåäñòàâëåíèÿ òåêñòîâîé èíôîðìàöèè è àâòîìàòèçàöèÿ ýò

– ñîçäàíèÿ ïðîôåññèîíàëüíûõ ñèñòåì è áàç çíàíèé;

– ïîääðæêè ïðîôåññèîíàëüíîé äåÿòåëüíîñòè ðàáîòíèêîâ ðàçëè÷íûõ îòðàñëåé;

– ïîâûøåíèÿ óðîâíÿ êîìïåòåíöèè ñïåöèàëèñòîâ çà ñ÷åò ïîëó÷åíèÿ âîçìîæíîñòè áûñòðîãî àíàëèçà è ïðåäñòàâëåíèÿ â óäîáíîé

– ïðîâåäåíèÿ ñèíòåçà òåêñòîâûõ äîêóìåíòîâ ñ ðàçëè÷íîé ñòåïåíüþ îáîáùåíèÿ èíôîðìàöèè;

– àâòîìàòèçàöèè ïðîöåññîâ ôîðìèðîâàíèÿ ñèñòåì îíòîëîãèé â òîé èëè èíîé ïðîôåññèîíàëüíîé îáëàñòè;

– ïðîâåäåíèÿ íàïðàâëåííîãî ïîèñêà è ôèëüòðàöèè òåêñòîâûõ äîêóìåíòîâ;

– àâòîìàòè÷åñêîãî ðåôåðèðîâàíèÿ òåêñòîâ äîêóìåíòîâ.

Summary

The paper considers the approaches associated with the vector representation of the textual information. The particularity of the approach under consideration is in the determination of the goal functions of separate sentences and representation of them in the form of some local vectors on which basis a global vector is built that determines the semantic component of the text on the whole. Various aspects of application of the proposed approach are considered.

Ñïèñîê ëèòåðàòóðû

Àðóòþíîâà 2005 Àðóòþíîâà, Í. Ä. Ïðåäèçâíåå è åãî ñìûñë / Í. Ä. Àðóòþíîâà. Ì. : ÓÐÑÑ, 2005.

Ìî÷åíîâ è äð. 2005 Ìî÷åíîâ, Ñ. Â. Ïðèìåíåíèå ñòàòèñòè÷åñêèõ ìåòîäîâ äëÿ ñåìàíòè÷åñêîãî àíàëèçà òåêñòà / Ñ. Â. Ìî÷åíîâ, À. Ì. Áë

Êàðàóëîâ è äð. 1982 Êàðàóëîâ, Þ. Í. Ðóññêèé ñåìàíòè÷åñêèé ñëîâàðü. Îïûò àâòîìàòè÷åñêîãî ïîñòðîåíèÿ òåçàóðóñà: îò ïîíÿòèÿ ê

Ðóáàøêèí è äð. 1998 Ðóáàøêèí, Â. Ø. Ñåìàíòè÷åñêèé (êîíöåïòóàëüíûé) ñëîâàðü äëÿ èíôîðìàöèîííûõ òåõíîëîãèé. ×. 1 / Â. Ø. Ðó-

Ñîêèðêî è äð. 2005 Ñîêèðêî, À. Ã. Ïðîåêò ÄÈÀËÈÍÃ, COM-îáúåêò Goldrml / À. Ã. Ñîêèðêî, Ä. Â. Ïàíêðàòîâ. Ì. : Äèàëîã, 2005.

Ôèíí 1999 Ôèíí Â. Ê. Î ðîëè ìàøèííîãî îáó÷åíèÿ â èíòåëëåêòóàëüíûõ ñèñòåìàõ // ÍÒÈ. Ñåð. 2. 1999. ¹ 12. Ñ. 1–3.

Salton 1989 G. Salton. Automatic Text Processing. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

Salton et al. 1994 G. Salton, J. Allan, and C. Buckley. Automatic structuring and retrieval of large text files. Communications of the ACM, 37(2), February 1994.

Todd et al. Todd A. Letsche and Michael W. Berry. Large-Scale Information Retrieval with Latent Semantic Indexing. URL: http://www.cs.utk.edu/ ~berry/sc95/sc95.html.