אחזור מידע

תרגום חופשי של הספר Information Retrieval

של C. J. van RIJSBERGEN

פרק 6 - אחזור הסתברותי

תרגום: עירן גישס, נועם עמישר וזאב גטנר

עריכה: עפר דרורי

הקדמה

עד כה בספר זה עשינו שימוש מועט בתורת ההסתברות בבניית מודל של איזו שהיא תת מערכת ב-IR. הסיבה לכך היא פשוט העובדה שעיקר העבודה ב- IR היא בלתי הסתברותית ורק לאחרונה נעשתה פריצת דרך משמעותית כלשהי עם שיטות הסתברותיות. ההיסטוריה של השימוש בשיטות הסתברותיות מגיעה עד תחילת שנות השישים, אך מאיזו שהיא סיבה הרעיונות הראשוניים לא השתרשו. בפרק זה אתאר שעטות אחזור, כלומר חוקי חיפוש ועצירה המבוססים על שיקולים הסתברותיים. בפרק 2 עסקתי בקטלוג אוטומטי מבוסס על מודל הסתברותי של התפלגות אסימוני מלים במסמך (טקסט): כאן אעסוק בהתפלגות של מונחי קטלוג בקובץ מסמכים היוצרים אוסף של קובץ. אסתמך בעיקר על ההנחה המוכרת שהתפלגות של מונחי קטלוג באוסף או בתת קבוצה שלו יגלה לנו משהו אודות החשיבות הצפויה של כל מסמך נתון.

יתכן שכדאי להזהיר את הקורא שחלק מהחומר בפרק זה הוא מתמטי. אולם אני מאמין שהמסגרת של האחזור הנדון בפרק זה היא גם אלגנטית וגם בפוטנציאל חזקה ביותר. חרף העובדה שהעבודה על כך הנה חדשנית למדי ולפיכך אחדים עשויים לחשוב שעליה לעמוד במבחן הזמן, לדעתי היא מייצגת את פריצת הדרך החשובה ביותר ב- IR בשנים האחרונות. לפיכך, ללא בושה, אעשה פרק זה תיאורטי היות והתיאוריה צריכה להיות מובנת היטב אם רוצים לעשות התקדמות כלשהי. ישנם מספר דרכים שונות להצגת התיאוריה הבסיסית. אני בחרתי להציג אותה בדרך כזו שהקשרים עם תחומים נוספים, כגון זיהוי תבנית, יעשו בקלות. יהיו לי דברים נוספים לומר אודות ניסוחים אחרים בהערות הביבליוגרפיות בסוף הפרק.

הכלי המתמטי הבסיסי בפרק זה הוא "משפט בייס". רוב המשוואות נובעות ישירות ממנו. למרות שהמתמטיקה הבסיסית עשויה בתחילה להראות מסובכת מעט התרגום הוא די פשוט. אז תנו לי לנסות ומייד לתת מעט תרגום למה שעתיד לבוא.

זכרו שהכלי הבסיסי שיש בידינו ע"מ להפריד את המסמכים הרלוונטיים מהבלתי רלוונטיים הוא פונקצית התאמה, הן אם אנו בסביבה מקובצת והן בסביבה בלתי מקובצת. הטעם בבחירת פונקצית התאמה מעולם לא הובהר, למעשה לרוב היא מבוססת על טענה אינטואיטיבית בהצלבה עם "OCKHAM'S RAZOR". בפרק זה אנסה להשתמש בתיאוריה הסתברותית לשם קביעה כיצד פונקצית התאמה צריכה להראות וכיצד לעשות בה שימוש. הטיעונים הם בעיקרם תיאורטיים אך לדעתי די נחרצים. הספק היחיד שנותר הוא על הקבילות של ההנחות שאנסה להציג בהסברי. המידע בו נעשה שימוש ע"מ לקבוע פונקצית התאמה נובע מהידע על התפלגות תנאי הקטלוג באוסף של מספר תת קבוצות שלה. אם היא מבוססת על מספר תת קבוצות של מסמכים אז תת קבוצה זו יכולה להיות מבוססת במספר דרכים: דגימה, קיבוץ או אחזור ניסיוני. המידע שנאסף משמש לקביעת ערכים של פרמטרים מסוימים הקשורים עם פונקצית ההתאמה. ברור שאם המידע מכיל מידע רלוונטי אז התהליך של הגדרת פונקצית ההתאמה יכול להיות מוזן ע"י מנגנון מישוב דומה לזה הנובע מ- ROCCHIO שתואר בפרק הקודם. בדרך זו הפרמטרים של פונקצית ההתאמה יכולים "להילמד". אנו נתרכז בפונקצית התאמה הנובעת ממידע רלוונטי.
יונח כי המסמכים מתוארים ע"י מאפייני מצב בינריים כלומר המצאות או היעדרות של תנאי קטלוג. זו אינה הגבלה על התיאוריה, בעקרון ההשלכה למאפיינים שרירותיים יכולה להימצא למרות שאין זה ברור שכדאי לעשות כן.

הערכה או חישוב של רלוונטיות

כשאנו מחפשים אוסף מסמכים אנו מנסים לאחזר מסמכים רלוונטיים מבלי לאחזר בלתי רלוונטיים. היות ואין עמנו נביא שיאמר לנו ללא טעות אילו מסמכים רלוונטיים ואילו בלתי רלוונטיים עלינו לעשות שימוש במידע שאינו מושלם ע"מ לנחש על כל מסמך שהוא אם הוא רלוונטי או בלתי רלוונטי. מבלי להיכנס לפרדוקס הפילוסופי הקשור לרלוונטיות אניח שאנו מסוגלים לנחש בלבד אודות רלוונטיות באמצעות מידע מתומצת על המסמך והיחסים שלו עם מסמכים אחרים. זו אינה הנחה מובנת במיוחד אם מאמינים שהדרך היחידה בה נקבעת רלוונטיות היא ע"י קריאת הטקסט כולו ע"י המשתמש. לכן דרך הגיונית לחישוב הניחוש שלנו היא לנסות ולהעריך לכל מסמך נתון את ההסתברות הרלוונטית שלו.

PQ(relevance/document)

הבה נניח (בעקבות רוברטסון [7]) :

1) הרלוונטיות של מסמך לבקשה היא בלתי תלויה במסמכים אחרים באוסף.

עם הנחה זו אנו יכולים להציג עיקרון, במונחי הסתברות של רלוונטיות, שניתן להשתמש במידע הסתברותי באופן אופטימלי באחזור. רוברטסון מייחס עיקרון זה ל-ו.ס. קופר למרות ש-מרון כבר הדגיש את האופטימליות שלו ב- 1964.

עקרון הדירוג ההסתברותי – אם תגובת מערכת שייכות אחזור לכל בקשה היא דרוג המסמכים באוסף בסדר יורד של רלוונטיות של הסתברות למשתמש שהגיש את הבקשה, בה ההסתברות מוערכת בדיוק רב ככל הניתן על בסיס העובדה האם המידע היה זמין למערכת למטרתנו, היעילות הכוללת של המערכת למשתמש שלה תהיה הטובה ביותר הניתנת להשגה על בסיס נתונים אלו.

כמובן עקרון זה מעלה שאלות רבות באשר לתקפות ההנחות. לדוגמא "היפותזת הצבר" שמסמכים הקשורים בינם נוטים להיות רלוונטיים לאותן בקשות, במפורש מניחה את ההפך מהנחה (1). GOFFMAN, גם כן, בעבודתו טרח רבות ע"מ לבסס הנחת תלות מפורשת. אני מצטט "אם מסמך X הוערך כרלוונטי לשאילתה S הרלוונטיות של מסמכים אחרים בקובץ X עשויה להיות מושפעת היות והערך של המידע הנמצא במסמכים אלו עשוי לגדול או להצטמצם כתוצאה מהמידע הנמצא במסמך X". כמו כן קיימת השאלה על הדרך בה נמדדת היעילות הכוללת.

רוברטסון, במאמרו מציג את עקרון הדירוג ההסתברותי למדידת יעילות במונחים של אחזור ונשורת. העיקרון גם נובע מהתיאוריה המתוארת בפרק זה. אך זהו אינו המקום להעלות את שאלות המחקר הללו, אולם, אני חושב כי זה הגיוני לאמץ את העיקרון כאחד שעליו ניתן לבנות מודל אחזור הסתברותי. מילת אזהרה, עקרון הדרוג ההסתברותי יראה נכון לשאילתה אחת בלבד. אין זה אומר שהביצועים על טווח של שאילתות יהיה אופטימלי, לבסס תוצאה מסוג זה יש להיות ספציפי על הדרך למצע את הביצועים על פני שאילתות.

עקרון הדרוג ההסתברותי מניח שביכולתנו לנחש F (רלוונטיות/מסמך), לא זו בלבד, הוא מניח שביכולתנו לעשות זאת בדיוק. זוהי הנחה בעייתית מאד והיא תעסיק אותנו בהמשך. הבעיה היא שאיננו יודעים אילו מסמכים רלוונטיים ואין אנו יודעים כמה ישנם כך שאין לנו דרך לחשב את F (רלוונטיות/מסמך). אך ביכולתנו באמצעות אחזור ניסיון, לנחש את F (רלוונטיות/מסמך) ובתקווה לשפר את הניחוש ע"י איטרציה. ע"מ לפשט את העניינים בדיון העוקב אניח שהסטטיסטיקות המתייחסות למסמכים רלוונטיים והבלתי רלוונטיים נגישות ואשתמש בהן ע"מ לבנות את המשוואות. אולם בכל זמן על הקורא להיות ער לעובדה שבכל מצב המידע היחסי צריך להיות מנוחש (או מוערך).נחזור לבעיה הנוכחית שהנה לחשב או להעריך את F (רלוונטיות/מסמך). לזאת נשתמש ב"משפט בייס" המייחס את ההסתברות הרלוונטית שאחרי להסתברות הרלוונטית שלפני וסבירות הרלוונטיות לאחר בחינת מסמך. לפני שנקפוץ לביטוי פורמלי של הנושא עלי להציג מספר סמלים שיפשטו את הדברים בהמשך.

מודל הסתברותי בסיסי

היות ואנו מניחים שכל מסמך מתואר ע״י תנאי קטלוג בהמצאות/היעדרות כל מסמך יכול להיות מיוצג ע״י וקטור בינארי

$X = (x1, x2, \dots , xn)$

כש- X1=0 או 1 מייצג העדות או המצאות של התנאי. אנו גם מניחים שישנם שני אירועים נבדלים

המסמך רלוונטי – W1
המסמך אינו רלוונטי – W2

אז במונחים של סמלי אלו, מה שברצוננו לחשב לכל מסמך זה את P(W/1X) ואולי את P(W/1X) כדי שנוכל להחליט איזה רלוונטי ואיזה אינו רלוונטי. זהו שינוי קל במטרה, מפשוט לייצר דרוג ברצוננו שהתיאוריה תאמר לנו כיצד לצמצם את הדרוג. לכן אנו מנסחים את השאלה כשאלת בחירה. כמובן שאין ביכולתנו להעריך את P(W/1X) ישירות לכן אנו חייבים למצוא דרך להעריך במושג של כמויות שאיננו יודעים עליהן.
משפט בייס מלמד אותנו על התפלגויות בדידות

$$P(w_i|x) = \frac{P(x|w_i)P(w_i)}{P(x)} \qquad i = 1,2$$

כאן P(wi) הוא עיקר הסתברות לרלוונטיות (i=1) או בלתי רלוונטיות, P(x/wi) פרופורציונלי למה שידוע כסבירות לרלוונטיות או בלתי רלוונטיות בהינתן X. במקרה הרציף זו תהיה פונקצית צפיפות ונרשום p(x/wi).
לבסוף

$$P(x) = \sum_{i=1}^{2} P(x/w_i)P(w_i),$$

שזו ההסתברות לצפות ב- X על בסיס רנדומלי בהינתן שעשוי להיות רלוונטי או בלתי רלוונטי. שוב זה ירשם בפונקצית צפיפות p(X) במקרה הרציף. למרות ש- P(x) (או p(x)) בעיקר יופיע כגורם נרמול (כלומר הבטחה ש- P(w1/x) + P(w2/x) = 1) שזו הפונקציה עליה אנו יודעים את הדברים הרבים ביותר, היא אינה דורשת מידע של רלוונטיות ע״מ להיות מוגדרת. לפני שאדון כיצד מעריכים את האגף הימני של משפט בייס אראה כיצד ההחלטה בעד או נגד רלוונטיות נעשית.

כלל ההחלטה בו אנו משתמשים ידוע היטב כ״חוק ההחלטה של בייס״. זהו
[P (w1/x) > P(w2/x) -> x is relevant, x is non-relevant] * D1

הביטוי D1 הוא קיצור של: השווה את P (w1/x) עם P (w2/x). אם הראשון גדול מהשני אזי החלט X רלוונטי, אחרת החלט X בלתי רלוונטי. המקרה בו P(w1/x) = P(w2/x) מטופל שרירותית ע״י קביעת בלתי רלוונטיות. הבסיס לחוק D1 הוא שהוא מצמצם את ההסתברות הממוצעת לטעות, הטעות שבקביעת מסמך רלוונטי כבלתי רלוונטי או להפך.
ע״מ לראות זאת שים לב שלכל X ההסתברות לטעות היא

$$P(error|x) = \begin{cases} P(w_2|x) & \text{if we decide } w_1 \\ P(w_1|x) & \text{if we decide } w_2 \end{cases}$$

המשמעות של [E -> p,q] היא שאם E נכון החלט P אחרת החלט Q.

במילים אחרות, מרגע שהחלטנו על דרך אחת (למשל רלוונטיות) אזי ההסתברות לטעות נתונה ע״י ההסתברות שהדרך ההפוכה היא הנכונה (כלומר בלתי רלוונטיות). ע״מ לעשות טעות זו קטנה ככל האפשר לכל X נתון עלינו לבחור ב- wi לו P (w1/x) הוא הגדול ביותר ובהתאם עבורו ההסתברות לטעות היא הקטנה ביותר. ע״מ לצמצם את ההסתברות הממוצעת לטעות עלינו למזער

$$P(error) = \sum_x P(error|x)P(x)$$

סכום זה ימוזער ע״י מזעור P (error/x) קטן ככל הניתן לכל X היות ו- P (error/x) ן P(x) תמיד חיוביים. מטרה זו מושגת ע״י חוק ההחלטה D1 שכעת מקבל הצדקה.

כמובן שטעות ממוצעת אינה הכמות היחידה שכדאי למזער. אם נקשר לכל סוג של טעות מחיר נוכל להסיר כלל החלטה שימזער ת הסיכון הכללי. הסיכון הכולל הוא ממוצע של תנאי הסיכון R(wi/x) שבפני עצמו מוגדר במושגים של פונקצית מחיר lij. כלומר lij הוא ההפסד הנגרם מההחלטה על wi כש- wj הוא הנכון. ההפסד הצפוי כשמחליטים wi נקרא סיכון משתנה ומוגדר

R (wi/x) - li1P(w1/x) + li2P(w2/x) i = 1, 2

הסיכון הכולל הוא סכום באותה דרך שההסתברות הממוצעת לטעות היתה R (wi/x) כעת ממלא תפקיד של P(wi/x). הסיכון הכללי ממוזער ע״י

R (w1/x) < R (w2/x) -> x is relevant, x is non-relevant] D2[

D1 ו- D2 יכולים להראות חליפיים תחת תנאים מסוימים. ראשית נשכתב D1 ע״י שימוש במשפט בייס בדרך בה יעשה בו שימוש חליפי

P(x/w1) P (w1) > P(x/w2) P(w2) -> x is relevant, x is non-relevant] D3[

שים לב ש- P(x) נעלם מהמשוואה היות שאין הוא משנה את התוצאה של ההחלטה. כעת ע״י שימוש בהגדרה R (wi/x) נקל להראות ש-

R (w1/x) < R (w2/x)] [[equivalence]] [(l21 - l11) P(x/w1) P(w1) > (l12 - l22) P( ]
P(x/w2) P(w2)[

כאשר פונקצית הפסד מיוחדת נבחרת

$$l_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

הרומז שלבחירה נכונה לא משתייך הפסד (הגיוני למדי) והפסד יחידה לטעות (פחות הגיוני) כאשר

R (w1/x) < R (w2/x) [[equivalence]] P(x/w1) P (w1) > P(x/w2) P(w2)[

המציג את שווה הערך ל- D2 ו- D3 ולפיכך ל- D1 ו- D2 תחת פונקצית הספד בינארית.

זה משלים את פיתוח חוק ההחלטה שישמש ע״מ להחליט על רלוונטיות או אי רלוונטיות, או להציג זאת באופן שונה, לשלוף או לא לשלוף. עד כה לא הושמו כל הגבלות על P(x/w1), לכן חוק ההחלטה הוא כללי למדי. הצגתי את הבעיה כבחירה בין שתי מחלקות ולפיכך התעלמתי מהבעיה של הדירוג. סיבה אחת לכך היא שכך הניתוח פשוט יותר והאחרת שברצוני שהניתוח יישאר למשתמש. בתוך המודל עד כה ניתן עדיין לדרג אך ערך החיתוך יהיה מפורש במונחים של הסתברויות קודמות ופונקציות מחיר. האופטימליות של עקרון הדרוג ההסתברותי נובע מהאופטימליות של חוק ההחלטה בכל נקודת חיתוך. עתה אבהיר את הצורה המדויקת של פונקצית ההסתברות בחוק ההחלטה.

פונקצית אחזור

החלק הקודם היה אבסטרקטי למדי והשאיר את הקשרים בין ההסתברויות השונות וה- IR פתוחים. למרות שזה הגיוני לרצות לחשב (רלוונטיות/מסמך)P אין זה כלל ברור כיצד לעשות זאת או האם ההמרה באמצעות משפט בייס היא הדרך הטובה ביותר. אף על פי כן נמשיך בהנחה ש- P(x/wi) היא הפונקציה הנכונה להנחה. פונקציה זו היא כמובן פונקצית הסתברות מצרפית והאינטראקציה בין רכיבי ס עשויה להיות מורכבת באופן שרירותי. ע"מ להסיק חוק החלטה שניתן לעבוד עמן יש להניח הנחה מקלה על P(x/wi). הדרך המתמטית הקונבנציונלית להפשטת P(x/wi) היא להניח כי מרכיבי המשתנים xi של X הם בלתי תלויים באופן סטוכסטי. באופן טכני זה מוביל להניח כי

P(x/wi) = P(x1/wi) P(x2/wi) ... P(xn/wi) A1

מאוחר יותר אראה כיצד ניתן יהיה לפשט הנחה קשיחה זו. בשלב זה נתעלם מהעובדה שהנחת התניות בלתי תלויות ב- W1 ו- W2 באופן נפרד הינה בעלת משמעות על התניות התלות על W1 ו- W2.

הבה עתה ניקח את הצורה המופשטת של P(x/wi) ונראה כיצד יראה חוק ההחלטה.

ראשית נגדיר מספר משתנים
pi = Prob (xi = 1/w1)

qi = Prob (xi = 1/w2)

במילים pi(qi) היא ההסתברות שאם מסמך הוא רלוונטי (בלתי רלוונטי) אזי תנאי קטלוג ה- i יהי נוכח. ההסתברות המתאימות להעדרות מחושבת ע"י חיסור מ- 1, כלומר pi = Prob (xi = 1 - 0/w1). פונקציית הסבירות שנכנסת ל- D3 תראה כעת

$$R(d|q) = \prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{1-x_i}$$

$$R(d|q) = \prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{1-x_i}$$

ע"מ להבין כיצד ביטויים אלו עובדים, על הקורא לבדוק ש-
P((0,1,1,0,0,1)/w1) = (1 - p1)p2p3(1 - p4)(1 - p5)p6. החלפת P(x/wi) ב- D3 ורישום, חוק ההחלטה יומר לפונקצית הפרדה לינארית

$$c_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

היכן שהקבועים ai ו- bi הם ברורים

$$g(x) = \sum_{i=1}^{n} (a_i x_i + b_i(1-x_i)) + a$$
$$= \sum_{i=1}^{n} c_i x_i + c$$

החשיבות שבכתיבה זו, מלבד פשטותה, היא שכל מסמך X ע"מ לחשב g(x) עלינו פשוט לחבר את המקדמים ci להם xi = 1. Ci לרוב נקרא משקל. רוסרטסון וספרק ג'ונס מכנים את ci כמשקל רלוונטיות וסלטון מכנה את exp(ci) כקבוע או משל. מכאן השם פונקצית שקילה ל- g(x).

הקבוע C שהנחנו כי הוא זהה לכל המסמכים X ישתנה משאילתא לשאילתא, אך ניתן לזהותו כחיתוך המיושם בפונקצית ההחזרה. החלק היחיד שניתן לשינוי בהתייחס לשאילתא נכונה היא פונקצית המחיר ושינוי זה הוא זה שיאפשר לנו לאחזר יותר או פחות מסמכים. ע"מ לראות זאת הבה נניח ש- l11 = l22 = 0 ושיש לנו בחירה בקביעת היחס l21/l11 ע"י בחירת ערך לחשיבות היחסית שאנו מיחסים לפספוס מסמך רלוונטי בהשוואה לשליפת מסמל בלתי רלוונטי. בדרך זו ביכולתנו לייצר דרוג, כל דרגה מתייחסת ליחס שונה של l21/l12.

הבה נעבור עתה לחלק האחר של g(x), כלומר ci וננסה לפרט זאת במונחים של טבלת ה- 'contingency' הקונבנציונלית.

	Relevant	Non-relevant	
$x_i = 1$	r	$n - r$	n
$x_i = 0$	$R - r$	$N - n - R + r$	$N - n$
	R	$N - R$	N

תהיה טבלה כזו לכל תנאי קטלוג. הראתי זאת לתנאי הקטלוג למרות שלא נעשה שימוש במקרא בתאים. אם יש לנו מידע מלא על המסמכים הרלוונטיים ובלתי רלוונטיים באוסף, אזי אנו נוכל להעריך את pi ע"י r/R ואת qi ע"י (n-r)/(N - R). לכן g(x) יכול להירשם

$$g(x) = \sum_{i=1}^{n} x_i \log \frac{r(R-r)}{(n-r)(N-n-R+r)} + C$$

זו למעשה פונקציה F4 בה עשו שימוש רוברטסון וספרק ג'ונס בניסויי "המבט לאחור" שערכו. לנוחות עתידית הבה נקבע

$$K_i(N,r,n,R) = \log \frac{r(R-r)}{(n-r)(N-n-R+r)}$$

יש מספר דרכים להכיל את Ki. הפירוש המעניין ביותר ל- Ki הוא לומר שהוא מודד את המידה בה התנאי ה- i יכול להבדיל בין המסמכים הרלוונטיים והבלתי רלוונטיים.

באופן טיפוסי ,המשקל, Ki(N,r,n,R) מוערך מטבלת ה- contingency בה N אינו המספר הכולל של המסמכים במערכת אלא הוא תת מערכת כלשהי שנבחרה במיוחד ע"מ לאפשר להעריך את Ki. מאוחר יותר אעשה שימוש בפרוש דלעיל של Ki ע"מ להציג פונקציה נוספת הדומה ל- Ki בכדי למדוד את כוח ההפרדה של תנאי הקטלוג.

מונחי הקטלוג אינם בלתי תלויים

למרות שנח מתמטית להניח שתנאי הקטלוג הם בלתי תלויים אין זה בהכרח ריאלי. ההתנגדות לחוסר תלות אינה חדשה. ב- 1964 ה.ה. ויליאמס הביע זאת כך "ההנחה של חוסר תלות של מילים במסמך נעשית בד"כ כחלק מנוחות מתמטית. בלי ההנחה רבים מהיחסים המתמטיים העוקבים לא יוכלו להיות מבוטאים. עמה רבות מההנחות צריכות להתקבל הזהירות יתר. רק בגלל שהמתמטיקה נהית די בלתי ניתנת למעקב כאשר מניחים תלות אנשים נוטים להניח חוסר תלות. אום "תלות היא נורמה יותר מאשר ההפך". בציטוט מהתאורתיקן ההסתברות הידוע דה-פינטי. לכן התהליך הנכון הוא להניח תלות ולאפשר לניתוח לפשט לחוסר תלות רק אם אכן יהיה כך. כאשר מדברים על תלות אנו מתכוונים לתלות סטוכטית. למידע IR תלות סטוכטית פשוט נמדדת ע"י פונקצית קורלציה או ע"י דרך דומה אחרת.

הנחת התלות יכולה להיות הכרחית כאשר אנו מנסים להניח (רלוונטיות/מסמך)P במונחים של P(x/wi), היות שהדיוק בה מוערכת ההסתברות האחרונה תשפיע ללא ספק על ביצועי האחזור. לכן משימתנו המיידית היא לעשות שימוש בתלות (קורלציה) בין תנאי האינדקס בכדי לשפר את הערכתנו של P(x/wi) עליה נשען חוק ההחלטה שלנו.

באופן כללי התלות יכולה להיות מורכבת בצורה הבאה

$$P(x) = P(x1)P(x2/x1)P(x3/x1,x2) \dots P(xn/x1,x2, \dots , xn - 1)$$

לכן, בכדי להתייחס לכל המשתנים התלויים נצטרך לבטא כל משתנה באמצעות קבוצה הולכת וגדלה של המשתנים הנוספים. בעקרון זה אפשרי אך זה כנראה יהיה מאוד לא יעיל חישובית, ובמקרים מסויימים אף בלתי אפשרי כאשר אין מספיק מידע לחשב מקרים בסדר גבוה. במקום זאת נאמץ שיטת הערכה ל P(x) שתיתן את מידע התלות המהותי. בצורה אינטואטיבית ניתן להסתכל בזאת כמישהו במסתכל על הפריסה הנ"ל ובוחר את המשתנה אם התלות הרבה ביותר. במילים אחרות אנו מחפשים תוצר מהצורה הבאה

$$P_t(x) = \prod_{i=1}^{n} P(x_i/x_{j(i)}) \qquad 0 \le j(i) < i \qquad A2$$

כאשר (m1, m2, ..., mn) מהווה פרמוטציה של השלמים 1 עד n ו j(.) היא פונקציה הממפה את i לשלמים קטנים מ i ו P(xi/xm0) שווה ל P(xi). להלן דוגמה לווקטור עם שישה מרכיבים x = (x1, ..., x6)

$$Pt(x) = P(x1)P(x2/x1)P(x3/x2)P(x4/x2)P(x5/x2)P\ (x6/x5)$$

שימו לב כמה דומה הנחה A2 להנחה הבלתי תלויה A1, ההבדל היחיד הוא בכך שב A2 לכל גורם קיים משתנה מתנה המזוהה עימו. בדוגמת הפרמוטציה (m1, m2, ..., m6)
הם (1,2, ..., 6) שזהו הסדר הטבעי, כמובן שהסיבה לכך שכתבנו את הפריסה כמו ב A2 היא להראות שהפרמוטציה (1,2, ..., 6) תיתן קירוב טוב. ברגע שניצור פרמוטציה זו ניתן לסווג את משתניה כבעלי סדר טבעי. הפרמוטציה ופונקציית j(.) המגדירות ביחד עץ תלות ואת Pt(x) המקביל נקראות התפלגות עץ התלות (מסדר ראשון). העץ המותאם לדוגמה עם ששת המשתנים מופיע באיור 6.1 .העץ מראה משתנה המופיע בכל צד של פעימת ההתניה ב P(./.). למרות שבחרנו לכתוב את הפונקציה Pt(x) בצורה ש xi הינו משתנה בלתי מותנה, ומכאן שורש העץ, למעשה כל עלה בעץ יכול לשמש כשורש העץ בתנאי שההתניה נעשית באופן רציף כלפי שורש העץ החדש.

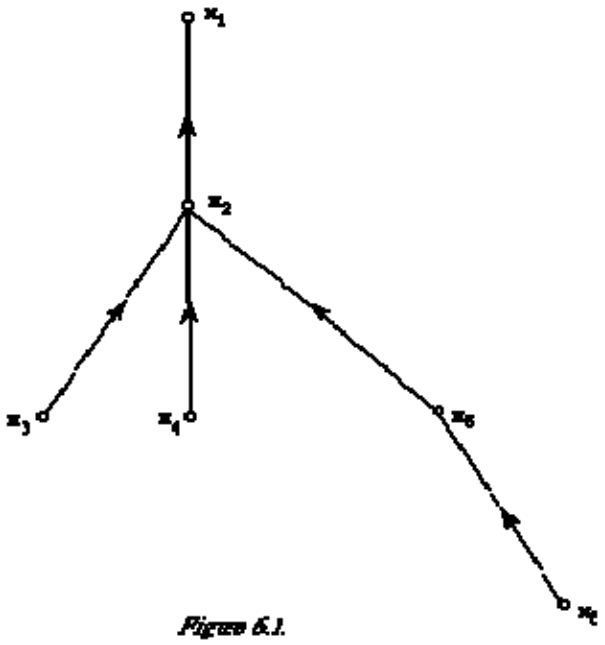

Figure 6.1.

$$P(x_{x_1}/x_{x_{X1}}) = P(x_{x_{X1}}/x_{x_1})\, P(x_{x_1})/P(x_{x_{X1}})$$

אם נפעיל את המשוואה על הקשר בין השורש X1 לבין הצאצא X2 בדוגמה השורש יעבור ל X2 והפריסה תשתנה להיות

Pt(x1, x2, ... x6) = P(x2)P(x1)/x2)P(x3/x2)P(x4/x2)P(x5/x2)P (x6/x5)

כמובן, בשביל לעמוד בחוק של מתן תוויות מחדש נחליף בין השמות '1' ו '2'. כל הפריסות המוחלפות בצורה זו הינם קירוב טוב ל P(x). ולכן העץ מייצג את קבוצת הפריסות המקבילות. ברור שקיימים מספר רב של עצי תלות, בעיית ההערכה שעלינו לפתור היא למצוא את העץ הטוב ביותר ; כלומר למצוא את הפרמוטציה הטובה ביותר ואת המיפוי j(.) הטוב ביותר.
החל מעכשיו נניח שנתינת התוויות נעשית בצורה xmi = xi.

בחירת עץ התלות הטוב ביותר

הבעיה העומדת בפנינו היא למצוא את פונקצית ההסתברות בצורת Pt(x) על קבוצת מסמכים שהיא הקירוב הטוב ביותר לפונקצית ההסתברות המשותפת P(x), וכמובן קירוב טוב יותר מזה הנובע מהנחה *A1. הקבוצה עליה הקירוב מוגדר יכולה להיות שרירותית ומוגדרת כאוסף המסמכים הרלוונטים או המסמכים הלא רלוונטים. לעת עתה לא נגדיר את הקבוצה. כשנבנה חוק החלטה הדומה ל D4 נאלץ למצוא קירוב ל P(x/w1) ו P(x/w2).

טיב הקירוב נמדד ע"י פונקציה המוגדרת היטב. אם P(x) ו Pa(x) הם שני הסתברויות דיסקרטיות אזי

$$I(P,P_a) = \sum_x P(x) \log \frac{P(x)}{P_a(x)}$$

הינו מדד לטיב הקירוב Pa(x). במונחים של פונקציה זו נרצה למצוא התפלגות של עצים תלויים Pt(x) כך ש I(P,Pt) מהווה מינימום, במילים אחרות למצוא עץ תלות מכלל העצים התלויים שיגרום ל I(P,Pt) להיות קטן ככל האפשר. במקרה וסטיית האינדקסים i ו j מאי תלות נמדדת באמצעות expected mutual information measure (EMIM) (ראה פרק 3)

$$I(x_i,x_j) = \sum_{x_i x_j} P(x_i,x_j) \log \frac{P(x_i,x_j)}{P(x_i)P(x_j)}$$

אזי הקירוב הכי טוב Pt(x), במובן שצמצום I(P,Pt), ניתן באמצעות עץ הפריסה המקסימלי (MST) (ראה פרק 3) על המשתנים x1 עד xn. עץ הפריסה נגזר מהגרף שצמתיו הם אינדקס 1 עד n וקצותיו משוקללים באמצעות I(xi,xj). ה MST הוא למעשה העץ ששקלול כל צמתיו

$$\sum_{i=1}^{n} I(x_i,x_{j(i)})$$

מהווה מקסימום. זוהי הגדרה מאוד כללית לעץ, אך הגדרה טובה יותר תהיה טכנית. הוכחה מדויקת של האופטימיזציה ניתן למצוא ב[13] Chow and Liu. אנו מעוניינים ביישום של מבנה העץ.

דרך אחת להסתכל על ה MST היא בכך שהוא כולל בתוכו את מרבית התלויות המשמעותיות בין המשתנים ומותנה בכך שסכום המשתנים צריך להיות המקסימום. לדוגמה באיור 6.1 הקשר בין המשתנים (צמתים x1 עד x6) הוכנס רק בגלל שהסכום

I(x1,x2) +I(x2,x3) + I(x2,x4) + I(x2,x5) + I(x5/x6)

הוא מקסימום. כל סכום יהיה קטן או שווה לסכום זה. שימו לב שאין זה אומר שכל משקל בפני עצמו המשויך לקצה יהיה גדול מקצה שאינו בעץ, למרות שכך זה ברוב במקרים.

ברגע שמצאנו את עץ התלויות קירוב ההתפלגות יכול להירשם בפורמט A2. מכאן ניתן לגזור פונקציה מאבחנת כמו שעשינו במקרה הבלתי תלוי

ti = Prob (xi = 1/xj(i) = 1)

ri = Prob (xi = 1/xj(i) = 0) ו r1 = Prob (x1 = 1)

P(xi /xj(i)) = [[ti^(xi)]*(1 - ti)^(1- xi)]* [xj^(i)]* [[ri^(xi)]*(1 - ri)^(1- xi)]^[1-xj(i)]

אז

$$\log P(x) = \sum_{i=1}^{n} [r_i \log r_i + (1-r_i)\log(1-r_i)] +$$

$$+ \sum_{i=1}^{n} \left[r_{j(i)} \log \frac{1-r_i}{1-t_i} + t_i r_{j(i)} \log \frac{t_i(1-r_i)}{(1-t_i)r_i} \right] + \text{constant}$$

זוהי פונקצית שקלול לא ליניארית הפשוטה יותר מזו הנגזרת מ A1 כאשר מניחים שהמשתנים בלתי תלויים, כאשר ti = ri. הקבוע הוא בעל אותה פרשנות במונחים של הסתברות אפריוריות ופונקצית הפסד. פונקצית ההחלטה השלמה היא

g(x) = log P(x/w1) - log P(x/w2)

שעכשיו מערבת חישובים בסדר גודל כפול מבמקרה הליניארי. רק הסכומים הכוללים xj(i) גורמים לפונקצית השקלול להיות שונה מבמקרה הליניארי, וזהו החלק המאפשר לקחת בחשבון ש xi תלוי ב xj. כשמשתמשים בפונקצית השקלול במסמך המכיל את xj(i) או את x(j)i וגם את xi יקבל את התרומה מחלק זה של פונקצית השקלול.

קל יותר לראות איך g(x) משלב משקלות שונות לתנאים שונים, אם מסתכלים במשקל שנתרם על ידי g(x) למסמך נתון x בשביל מצבים שונים של הזוג xi xj(i). כאשר xi = 1 וגם xj(i) = 0 המשקל הנתרם הוא

$$\log \frac{\text{Prob}(x_i = 1 | x_{j(i)} = 0) \wedge e_1)}{\text{Prob}(x_i = 1 | x_{j(i)} = 0) \wedge e_2)}$$

ובאופן דומה בשביל שלושת המצבים xi וגם xj(i)

כעת ראינו עד כמה פשוטה פונקצית השקלול הלא ליניארית. לדוגמה בהינתן מסמך ש i מופיע אבל j(i) לא מופיע, אז המשקל הנתרם באמצעות g(x) מבוסס על היחס בין שתי הסתברויות. הראשונה היא המופע של i בקבוצת מסמכים רלוונטים כשנתון ש j(i) לא מופיע, השניה היא האנלוגיה על קבוצת מסמכים לא רלוונטית. על בסיס יחס זה אנו מחליטים אם יש מספיק נתונים בשביל להציב x עבור מסמכים רלוונטים או לא רלוונטים. חשוב לזכור בשלב זה שהנתונים בשביל לבצע את ההצבה בדרך כלל מתבססת על ההערכה של זוג הסתברויות.

הערכת משתנים

השימוש בפונקצית שקלול מהסוג הנ״ל באחזור דורש הערכה של פרמטרים מקושרים. כעת נבצע הערכה של ti ו ri במקרה הלא ליניארי, המקרה הליניארי יוצג לאחר מכן. ניתן כעת דוגמה של תהליך ההערכה המשתמש במקסימום פשטני באמצעות דמיון. הבסיס להערכות הוא הטבלה הבאה

	$x_i = 1$	$x_i = 0$	
$s_{jk} = 1$	[1]	[2]	[7]
$s_{jk} = 0$	[3]	[4]	[8]
	[5]	[6]	[9]

כאן אימצנו את הסימון לתאים כך ש [x] משמעות מספר המופעים בתא x. אם נתעלם מכך שהקבוצה מבוססת טבלה; הערכותינו יהי כדלקמן:

$$P(s_i = 1 | x_{jk} = 1) = r_i = \frac{[1]}{[7]}$$

$$P(s_i = 1 | x_{jk} = 0) = r_i = \frac{[3]}{[8]}$$

במקרה הכללי יהיו בידינו שתי טבלאות מסוג זה היוצרות את הפונקציה g(x), האחת בשביל להעריך פרמטרים המקושרים עם P(x/w1) והשניה בעבור פרמטרים המקושרים עם P(x/w2).
בגבול יהיה לנו מידע מלא איזה מסמכים שנאספו רלוונטים ואיזה לא. אם היינו מחשבים את ההערכות למקרה הגבולי היינו מקבלים את הגבול העליון למקרה האחזור במקרה זה. באופן מציאותי יותר יהיה בידינו מדגם של מסמכים, סביר להניח קטן (לא בהכרח מקרי), כך שבעבור כל מסמך מידת הרלוונטיות הייתה ידועה. המקבץ הקטן יהווה מקור נתונים לכל טבלת 2 על 2 שנרצה לבנות. ההערכות לכן יהיו משוחדות בצורה בלתי נמנעת.
הערכות הנ״ל הינן דוגמאות להערכות בנקודה. קיימות מספר דרכים בשביל לקבל חוקיות המתאימה להערכה בנקודה. לצערנו חוק ההערכה המוצלח ביותר הינו עדיין בעיה פתוחה [14].
למעשה סטטיסטיקנים מאמינים שאין לבצע כלל הערכה בנקודה [15]. למרות זאת באחזור מידע קשה לראות כיצד נמנעים מהערכה בנקודה. בכל חוק הערכה בנקודה נעשות מספר הנחות שרירותיות. למזלנו באחזור מידע קיים סיכוי לתקן הנחות אלו באמצעות מידע הנאסף ממערכות אחזור, ובכך מסירים חלק מהשרירותיות.
שתי הנחות מרכזיות נעשות כאשר יוצרים חוקי ההערכה על פי תיאוריית בייס:

(1) הצורה של ההתפלגות הקודמת במרחב המשתנים וגם

(2) הצורה של פונקצית ההפסד המשמשת למדידת הטעות הנעשית בעת הערכת המשתנה.

ברגע ששתי הנחות אלו קיימות ומוגדרות היטב על ידי הגדרתם כהתפלגות ופונקצית הפסד, אז, ביחד עם עיקרון בייס המחפש לצמצם את הטעות בהינתן התצפית, ניתן להסיק מספר חוקי ההערכה. הספרות הסטטיסטית לא נותנת העדפות לחוקים מסוימים. (לפרטים ניתן להסתכל ב van Rijsbergen[2]). החוקים החשובים של הערכת a ביחס ל p כולם באים בצורה

$$\hat{p} = \frac{x + a}{n + a + b}$$

כאשר x הוא מספר ההצלחות מתוך n ניסיונות, a ו b הם משתנים המוכתבים על ידי השילוב בין פונקצית הקדימות וההפסד. כך יש בידינו קבוצה שלמה של חוקי ההערכה. לדוגמה כאשר a=b=0 יש לנו את ההערכה x/n, וכאשר a=b=0.5 יש לנו את החוק של Good[16]. החוק האחרון הוא למעשה החוק ש Robertson and Sparck Jones[1] השתמשו בו בהערכותיהם. כל מצב של a ו b ניתן להצדיק בתנאים של סבירות תוצאות התפלגות אפריורית.
מאחר ומה שסביר בעיני אדם אחד אינו סביר בעיני אדם שני, ההחלטה הסופית צריכה להתבסס על הביצועים בניסויים. למזלנו באחזור מידע אנו במצב ייחודי לבצע ניסויים אלו.
סיבה חשובה אחת כדי שחוקי הערכה יהיו שונים מהצורה הפשטנית x/n, היא בכך שזה אינו סביר בעבור מדגמים קטנים. לדוגמה מדגם עם דגימה אחת (n=1) ותוצאה (x=0) או (x=1) שיגרום לתוצאה בעבור p (p=0) או (p=1). ברור שזהו מגוחך, מאחר וברוב המקרים אנחנו יודעים בהסתברות גבוהה ש $0 < p < 1$.

בשביל להתגבר על קושי זה אנו יכולים לשלב מידע מוקדם בהתפלגות על הערכים האפשרים אותם אנו מנסים להעריך. ברגע שנקבל את הפיסבילוית של זה ונגדיר דרך למדוד את טעות ההערכה, עיקרון בייס (או עקרונות אחרים) לרוב יובילו לתוצאה שונה מ x/n.
לא נוסיף יותר בנושא זה, קיים פירוט רב יותר בספרות סטטיסטית.

סיכום

בשלב זה נסכם בשפה פשוטה. הנקודה הראשונה היא שעד כה ניסינו להעריך את P(relevance/document) כלומר ההסתברות שמסמך יהיה רלוונטי. למרות שקל לכתוב משפט זה לא ברור שיש לו משמעות. רלוונטיות של מסמך זהו מונח שאינו מוגדר היטב, יש שיטענו שזה יכול להיות 1 או 0. אחרים יטענו שניתן להסיק זאת על ידי ההסתברות P(x/relevance) לפי נוסחת בייס. לא נטיל ספק בשימוש בנוסחת בייס השאלה אם ל P(relevance/x) יש משמעות בהקשר של אחזור מידע. נניח שכן כדי שנוכל לקשר בין P(relevance/x) לבין התפלגות המידע בעבור אינדקסים.

גישה זו הייתה חסרת משמעות אלמלא היינו מניחים שההתפלגות להרבה אינדקסים למסמכים רלוונטים שונה מזו למסמכים לא רלוונטים. אם היינו מניחים את ההפך אזי P(relevance/document) יתפלג כמו P(relevance), קבוע לכל המסמכים ומכאן לא מסוגל להחליט מי מהמסמכים הוא חסר ערך לאחזור. מכאן אנו מניחים שמידע עקיף נובע מההתפלגות המשותפת של אינדקסים מעל שתי קבוצות שיעזור לנו להבחין בינם. ברגע שנקבל השקפה זו אנו מחויבים לניסוח הנגזר מן הנאמר. המחויבות היא שאנו חייבים לנחש את P(relevance/document) בדיוק מרבי, או לחילופין לנחש את P(document/relevance) וגם את P(relevance) באמצעות התפלגות המידע הנובעת מהמסמך.

ההסבר במונחי דירוג במקום אבחנה פשוטה הוא טריוויאלי: ההפחתה הנובעת מהקבוע g(x) בהדרגתיות קטנה ובאמצעות זאת מגדילה את מספר המסמכים המאוחזרים. תוצאות הדירוג הם אופטימאליות עקב העובדה שעבור כל הפחתה אנו ממזערים את הסיכון הכללי. צריך להתייחס לכך בזהירות כי זה מניח שיש בידינו את P(x/wi) בצורה נכונה והערכות שלנו הם הטובות ביותר שאפשר, לא סביר שיתממש בניסוי.

אם נסכים שהמשתמש יקבע את גודל ההפחתה לאחר האחזור אזי הצורך בתיאוריה בנושא ההפחתה בטל. ההשלכה הנובעת מכך שבמקום לעבוד אם היחס

$$\frac{P(x/relevance)}{P(x/non\text{-}relevance)}$$

נעבוד אם היחס

$$\frac{P(x/relevance)}{P(x)}$$

במקרה האחרון איננו מסתכלים על בעיית האחזור כאבחנה בין מסמכים רלוונטיים ללא רלוונטיים, במקום אנו מנסים לחשב את ההסתברות P(relevance/x) לכל מסמך x ולהציג גודל זה בסדר יורד למשתמש. בכל צורה שנסתכל בזה עדיין נדרש הערכה של שתי פונקציות הסתברות משותפת.

חוקי ההחלטה הנובעים מהנ"ל מבוטאים בצורה של P(x/wi). מכאן אדם יכול להסיק שהערכה של הסתברויות אלו הכרחית לביצועי האחזור, וכמובן העובדה שרק ניתן להעריך הסתברויות אלו היא הסבר לחלק מחוסר האופטימאליות של הביצועים. בכדי להעריך צריך להניח מספר דברים לגבי הצורה של P(x/wi). הנחה ברורה היא להניח חוסר תלות סטוכסטי לגבי מרכיבי x. באופן כללי זה לא ריאלי מאחר וטבע פונקצית האחזור שאינדקסים יהיו תלויים אחד בשני. נצטט מאמר מוקדם של "מרון" בנקודה זו: "בשביל לעשות זאת אדם יצטרך ליצוא תוכנת מחשב לבצע ניתוח אנליטי של אינדקסים כך שהמחשב "יידע" מי מהתנאים קרוב ביותר אל תנאי אחר ויכול להצביע כל הכיוון הוודאי ביותר שבו ניתן להגדיל בקשה". לכן גישה ריאלית יותר היא להניח תלות כלשהי בין התנאים כאשר מעריכים את P(x/w1) ואת P(x/w2) או P(x).

כעת נמשיך ונדון בדרכים להשתמש במודל זה של אחזור ובוזמנית נדון בבעיות הצצות. בהתחלה לא נשנה כמעט את המודל אבל לאחר מכן הדון בדרכים מעשיות למימוש שלא בהכרח ייצמדו להנחות על פיהם בנינו מודל זה. באופן טבעי ההצדקות לכך טמונות בניסויים שרבים עדיין נותר לבצע [17]. נתחיל בהסבר לשינוי קטן הנובע בצורך לצמצם את ממדי הבעיה.

מגרעת הממדים

בהסקת חוקי ההחלטה הנחנו שמסמך מוצג באמצעות ווקטור בעל n ממדים כאשר n הוא גודל האינדקסים. בדרך כלל n יהיה גדול למדי, ולכן ממדי וקטורי המסמך יהיה לרוב גדול ממספר הדגימות המשמש להעריך את הפרמטרים בפונקצית ההחלטה. דבר זה יוביל לבעיות שצוינו באופן מתמיד בספרות בנושא זיהוי דפוסים. למרות שניתוח הבעיה בזיהוי דפוסים רלוונטי לאחזור מידע גם כן, הפתרון אינו ישים מיידית. בזיהוי דפוסים הבעיה היא: בהינתן מספר הדגימות ששימשו "לאמן" את פונקצית ההחלטה (פונקצית השקלול שלנו), האם קיים מספר אופטימאלי של מדידות שניתן ליצור מדפוס בלתי ידוע כך שממוצע ההסתברויות של השמה נכונה ניתן למקסמו? במקרה שלנו בכמה אינדקסים ניתן להשתמש כדי לקבוע רלוונטיות. היוז[18] מראה שבעבור מבנה הסתברותי כללי מספר המדידות הוא קטן באופן מפתיע למרות שדגימות בגודל סביר משמשות בכדי "לאמן" את פונקצית ההחלטה.

באופן אידיאלי נרצה לבחור תת קבוצה (קטנה) של אינדקסים שבעבורם פונקצית השקלול g() תוגבל ובכך תמקסם את ממוצע ההסתברויות של השמה נכונה. בזיהוי דפוסים יש טכניקות מסובכות לעשות זאת לבעיות מקבילות. באחזור מידע התמזל מזלנו וקיימת דרך טבעית שבאמצעותה ניתן לצמצם את ממדי הבעיה. אנו מקבלים שתנאי השאלה הם מדריך טוב לתכונות המוצלחות ביותר ביישום של פונקצית g() בשביל להחליט בין רלוונטיות ואי רלוונטיות של מסמכים. לכן במקום לחשב את פונקצית השקלול לכל התנאים אנו מגבילים את g() לתנאים המצוינים בשאלה ואפשרי שגם לתנאים דומים. המשמעות של הנ"ל היא שבזמן תהליך האחזור כל המסמכים מוקרנים ממרחב רב ממדי לתת מרחב הנפרש באמצעות מספר תנאים קטן יותר.

פרטים חישוביים

נעבור כעת לחלקים המעשיים יותר בחישוב g(x) לכל x כאשר מניחים שהמשתנים xi תלויים סטוכסטית. המטרה המרכזית של חלק זה היא להוכיח שהחישובים הנם פיזיביליים. הדרך הבהירה ביותר לעשות זאת היא לחשב עבור כל אובייקט את ה EMIM את ה MST ואת ה g(.) בנפרד ובסדר זה.

1. חישוב ה EMIM

ניתן לפשט את חישוב ה EMIM וכן ניתן להעריך אותו וכך להפחית את זמן החישוב. נבחן תחילה את הדרך הראשונה:

כשמחשבים את l(xi,xj) לצורך בניית ה MSTאנו צריכים לדעת רק את סדר הדירוג עבור כל ה l(xi,xj) . הערכים עצמם אינם משנים. לכן ניתן להשתמש במידע על ההסתברויות בטבלה הבאה :

	$x_i = 1$	$x_i = 0$	
$x_j = 1$	[1]	[2]	[7]
$x_j = 0$	[3]	[4]	[8]
	[5]	[6]	[9]

במקרה זה תהיה הפונקציה I(xi,xj) מונוטונית לגמרי עם הפונקציה:

$$[1]\log\frac{[1]}{[5][7]} + [2]\log\frac{[2]}{[6][7]} + [3]\log\frac{[3]}{[5][8]} + [4]\log\frac{[4]}{[6][8]}$$

וזוהי אכן דרך פשוטה לחישוב ה EMIM . נחשוב כעת על מקרה שבו אנו רוצים לחשב את p(x) .במקרה זה, ה MST מבוסס על מידע המופיע מספר פעמים והנגזר מכל האוסף. ברגע שיש לנו את זה ([1]) ואנו יודעים את מספר המסמכים בקובץ ([9]) אז כל קובץ הפוך יכיל את יתר המידע על התדירות שאנו צריכים כדי למלא את יתר התאים בטבלה. כלומר, בעזרת ([5]) ו ([7]) הנתונים על ידי הקובץ ההפוך אנו יכולים להסיק את תוצאתם של [2] [3] [4] [6] ו [8].

ישנה בעיה של טיפול בערכי 0 בתאים 1-4 והיא נפתרת אם קובעים שהתוצאה של 0log0 שווה ל אפס. שאר התאים אינם יכולים לקבל ערכי אפס.

כעת נדון בדרך השנייה שבה מעריכים את ה EMIM מרון וקונס השתמשו בעבודתם המוקדמת בנוסחה:

$$d(xi,xj) = P(xi = 1, xj = 1) - P(xi = 1)\,P(xj = 1) \quad (*)$$

כדי למדוד את הסטייה מעצמאות עבור שני אברי אינדקס I,j . לבד מה log זוהי הדרך הראשונה של הרחבת ה EMIM. אם מחשבים את ה MST (עץ תלות) לפי הנוסחה הקודמת, תוצאות החישוב של p(x|wi) לא יחיו מושלמות אבל ההערכה תהיה קרובה מספיק מחישוב ותחסוך זמן רב לעומת חישוב מלא של ה EMIM. באופן דומה, לוי השתמש בנוסחה:

$$\log\frac{P(x_i = 1, x_j = 1)}{P(x_i = 1)\,P(x_j = 1)}$$

להערכת ה EMIM. ללא ספק, ישנן דרכים נוספות להערכת ה EMIM אך על השאלה אם ניתן להשתמש בהן למציאת עץ תלות שבעזרתו ניתן להגיע להערכה טובה של פונקצית ההסתברות המשותפת ניתן לענות לאחר ניסויים אמפיריים.

2. חישוב ה MST :

ישנם מספר אלגוריתמים לחישוב ה MST . כנראה שהטוב שבהם הוא האלגוריתם החדש של וויטני שהסיבוכיות שלו היא o(n[2]) כאשר n הוא מספר איברי האינדקס בעץ התלות. זהו אינו מכשול לשימוש באלגוריתם עבור קבוצות גדולות היות וקל לחלק את המידע באמצעות טכניקות שנידונו קודם ואחר כך, ניתן לבנות את העץ על ידי הפעלת האלגוריתם על כל אשקול של איברי אינדקס . דבר זה מוריד את הסיבוכיות ל o(k[2]) כאשר k קטן בצורה משמעותית מ n.

לפי עקרונות אלו, הוכיחו בנטלי ופרידמן ששימוש במרחב הנדסי שבו איברי האינדקס הם נקודות במרחב, הסיבוכיות של האלגוריתם לחישוב ה MST. יורדת ל o(n*log(n) בקרוב ויותר מכך ניתן לבנות עץ מפתח שהוא כמעט MST שגם הסיבוכיות שלו היא o(n*log(n).

ישנו חוסר יעלות גדול, בכך שכל n*(n-1)/2 האסוציאציות מחושבות ורק בחלק קטן מהן יש משמעות בהקשר שהן אינן ערך אפס וניתן להשתמש בהן בעץ המפתח. רוב האסוציאציות מקבלות ערכי אפס וניתן להתעלם מהן, אך לרוע המזל ניתן לדעת איזו אסוציאציה תקבל ערך אפס רק לאחר חישוב. קרופט, גילה לאחרונה להתעלם מאסוציאציות מבלי לחשבן תחילה. דרך זאת מניחה שהקובץ והצורה ההפוכה שלו קיימים ולכן יתכן שבאלגוריתם של קרופט נדרש זמן להפוך את הקובץ.

3. חישוב g(x)

יש להדגיש שבמקרים לא ליניאריים, הערכת הפרמטרים של g(x) תגרום באופן אידיאלי ל MST שונה עבור כל p(x|w1) ו p(x|w2) . כמובן ישנו מידע מלא על הופעת איברי האינדקס בקבוצות ה שייכות\חוסר שייכות במצב הניסויי. חישוב g(x) באמצאות מידע מלא עשוי להיות נחוץ כשגוזרים את הגבולות העליונים של יעילות האחזור לפי המודל, כפי שנעשה במקרה העצמאי אצל רובר טסון, ספארק וג'ונס. במקרה שבו אין מסמכים רלוונטיים שידועים מראש, יהיה צורך להשתמש בטכניקה של משוב רלוונטיות כדי להעריך את הפרמטרים ואז הביצועים ישאפו לגבול העליון. עובדה זאת, מובטחת לפי תיאוריית השאיפה שנדונה בעמוד 106 בפרק 5.

ישנם מספר דרכים לפרש את המודל לגבי g(x) והדבר תלוי בשאלה האם רוצים לקבוע את ההסתברות אפריורית או אפסוטיורית. במקרה הראשון יש לקבוע את ה MST הן עבור איברי האינדקס במסמכים הרלוונטיים והן במסמכים הלא רלוונטיים. היות וניתן לעשות זאת רק עם מידע מדגמי ולא מלא, עץ התלות יהיה רחוק מן האופטימאלי. דרך היוריסטית להתגבר על הבעיה היא לבנות עץ תלות עבור כל האוסף. מבנה העץ צפוי להיות במקרה זה זהה למבנה של שני עצי התלות המבוססים על המסמכים הרלוונטיים והלא רלוונטיים . p(x|w1) ו p(x|w2) מחושבים על ידי חישוב ההסתברות המותנית עבור הצמתים בעץ התלות. עד כמה חישוב זה מדויק, ניתן לקבוע באמצעות ניסויים בלבד.

אם מניחים שההסתברויות נקבעות אפוסטריורית, הרי ניתן לדרג את המסמכים לפי p(w1|x) ולהשאיר למשתמש את ההחלטה מתי הוא ראה מספיק. כלומר, אנו משתמשים בנוסחה:

$$P(w_1 \mid x) = \frac{P(x \mid w_1)\, P(w_1)}{P(x)}$$

להערכת ההסתברות של רלוונטיות עבור כל מסמך X.כאן אנו זקוקים להערכה בלבד של p(x|w1) ולאחר חישוב p(x) אנו משתמשים בעץ מפתח עבור כל האוסף מבלי להתייחס למידע שייכות. לגישה זאת מספר יתרונות. יתרון אחד הוא, שאם יש הנחת עצמאות עבור כל האוסף הרי יש קונסיסטנטיות עם הנחת העצמאות עבור המסמכים הרלוונטיים . הנחות אלו יגבירו את החישובים בצורה משמעותית. למרות שהנחת העצמאות עבור W1 איננה סבירה, היא עשויה להיות כורך המציאות, היות וקשה למדוד את התלות המדויקת בניסויים או במדגמים.

דרך אלטרנטיבית לשימוש בעץ תלות (היפוטיזת האסוציאציות)

לפי הטיעונים בפסקה הקודמת ניתן להסיק שעץ התלות היחיד האפשרי הוא עץ תלות עבור כל האוסף. נמשיך הלאה לפי קו זה. בניית עץ תלות עבור איברי האינדקס מבלי להשתמש במידע על תלות בניהם דומה לבניית טבלת סיווג לאיברי אינדקס. בפרק 3 ציינתי את הקשר בין הMST והקשר היחיד והוכחתי שהאחד אינו כה שונה מהשני. עובדה זאת מובילה לרעיון ששימוש בעץ תלות דומה לשימוש בקיבוץ איברים.

הרעיון של קיבוץ איברים הוסבר בפרק 2. ניתן לסכם זאת באמירה שאפשר ליישם מספר אסטרטגיות של הוספת ומחיקת איברים לפי השיטה של קיבוץ איברים. אם נתעלם לרגע מהמחיקה, ברור כיצד ניתן להשתמש בעץ התלות כדי להוסיף איברים או להרחבת השאילתה. הסיבה לכך ניתנה ב1964 ע"י מרון ששאל: "כיצד ניתן להגדיל את ההסתברות לאחזור מספר מסמכים המכילים מידע ייחודי". דרך ברורה מאליה היא להרחיב את הבקשה המקורית באיברי אינדקס בעלי משמעות קרובה לאינדקסים בבקשה המקורית. ההנחה כאן היא שמשמעות קשורה יכולה להתגלות ע"י אסוציאציה סטטיסטית. לכן אני מציע שבהינתן שאילתה שבה הקריטריונים אינם ברורים, אנו משתמשים בעץ התלות כדי לגלות אילו איברים נוספים הקשורים באיברים בשאילתה עשויים לעזור לנו במציאת מסמכים רלוונטיים. כלומר אני טוען שאיברי אינדקס הקשורים ישירות לאיבר מתוך שאילתה בעץ תלות, עשויים לעזור באחזור. למעשה ניסחתי מחדש את ההיפוטיזה עליה מבוססת תיאורית קיבוץ האיברים. ברצוני כעת לנסח זאת פורמאלית ולקרוא לכך היפוטיזת האסוציאציות: אם איבר אינדקס כלשהו יכול להועיל לסינון מסמכים רלוונטיים ומסמכים לא רלוונטיים, אז איברי אינדקס בעלי אסוציאציות לאיבר זה יועילו ככל הנראה גם הם.

אנו מפרשים היפוטיזה זאת בכך שאם איבר בשאילתה נבחר ע"י המשתמש, אזי הוא מסנן טוב בין מסמכים רלוונטיים ולא רלוונטיים ולכן אנו מעוניינים באסוציאציות הקרובות שלו. ההיפוטיזה אינה דנה בדרכים שבהם ניתן למדוד את האסוציאציה בין איברי אינדקס אך בפרק זה דנתי בשימוש ב EMIM. היא גם לא דנה בכימות הסינון, בכך אדון בפסקה הבאה. במידה מסוימת בת זוג של תיאוריית קיבוץ האיברים וניתן לבחון אותן באותה דרך.

כוח הסינון של מונח אינדקס:

בעמוד 120 הגדרתי:

$$K_i(N,r,n,R) = \log \frac{r(R-r)}{(n-r)\,|\,(N-n-R+r)}$$

ולמעשה הערתי שנוסחה זאת מודדת את הכוח של איבר i לסנן בין מסמכים רלוונטיים ולא רלוונטיים. המשקולות בפונקציה נגזרים מהנחת העצמאות. אם נתעלם לרגע מכך שמשקלים אלו הם תוצאה של מודל מסוים, ולחלופין נתייחס לכוח הסינון של איבר אינדקס כדבר העומד בפני עצמו. זהו בבדאי אינו דבר קל לעשותו. סלטון בעבודתו חיפש דרכים אפקטיביות למדוד את כוח הסינון של איבר אינדקס. זה ניראה הגיוני להצמיד לכל איבר אינדקס הנכנס לתהליך אחזור, משקל המתייחס לכוח הסינון שלו. שימוש ב Ki כדרך למדוד כוח זה בעייתי במקצת היות והוא לא מוגדר כאשר ארגומנטים של פונקצית הלוגריתם הופכים לאפס. לכן אנו מחפשים דרך חלקה יותר למדידת כוח הסינון. הפונקציה עליה אני עומד להמליץ היא אכן חלקה יותר ומאפשרת לי להסיק מסקנה כללית בעלת חשיבות רבה. יש להדגיש עם זאת שהפונקציה המוצעת היא רק אפשרות אחת, אך כל פונקציה מתאימה תהייה דומה לזאת המוצעת כאן.

במקום להשתמש ב Ki אני מציע להשתמש ברדיוס המידע כפי שמוגדר בפרק 3 בעמוד 42, כדרך למדידת כוח הסינון של איבר אינדקס. זהו קרוב משפחה של "אמצעי המידע המשותף הצפוי" שנשתמש בו בהמשך. נשתמש ב U וב V בתור משקולות חיוביות כאשר 1=V+U ובסימונים הרגילים של פונקציות הסתברות. אנו יכולים לנסח את רדיוס המידע כך:

$$\begin{aligned}
&uP(x_i=1|w_1)\log\frac{P(x_i=1|w_1)}{uP(x_i=1|w_1)+vP(x_i=1|w_2)} + \\
&vP(x_i=1|w_2)\log\frac{P(x_i=1|w_2)}{uP(x_i=1|w_1)+vP(x_i=1|w_2)} + \\
&uP(x_i=0|w_1)\log\frac{P(x_i=0|w_1)}{uP(x_i=0|w_1)+vP(x_i=0|w_2)} + \\
&vP(x_i=0|w_2)\log\frac{P(x_i=0|w_2)}{uP(x_i=0|w_2)+vP(x_i=0|w_2)}
\end{aligned}$$

הפירוש המעניין של רדיוס המידע מוצג בצורה בהירה יותר במונחים של פונקציות הסתברות רציפות. במקום להשתמש בצפיפויות של p(,|w1) ו p(,|w2 אשתמש בהסתברויות התואמות של U1 ו U2 . נגדיר תחילה את הממוצע של שתי הסתעפויות ישירות:

R (u1, u2/*v*) = *uI* (u1/*v*) +*vI* (u2/*v*)

l(ui|v) מודד את כמות המידע שמרוויחים מהידיעה שיש לדחות את V לטובת Ui . כעת רדיוס

$$R(\mu_1,\mu_2|v) \quad \frac{uu}{v}$$

המידע הוא המינימום של:

וכך מורידים את ה V השרירותי. למעשה, מסתבר שהמינימום מושג כאשר:

v = *u* u1 + *v* u2

זהו למעשה ממוצע של שני הערכים שיש לסנן. אם אנו מקבלים את V ו U כהסתברויות פריוריות אז Vמובע בצפיפות:

p(***x***) = *p*(***x***/*w*1) *P*(*w*1) + *p*(***x***/*w*2) *P*(*w*2)

המוגדרות עבור כל האוסף ללא התייחסות לרלוונטיות. בערך זה אנו בטוחים אך את הערכים U1 ו U2 יש לנחש. לכן זה הגיוני שבעת מדידת ההפרש בין U1 ו U2 V ימזג את כל המידע הקיים. רדיוס המידע בדיוק עושה זאת.

ישנה בעיה טכנית אחת בשימוש ברדיוס המידע או בכל דרך מדידת סינון אחרת המבוססת על ארבעת התאים בטבלת השכיחות. הבעיה היא ששיטות אלו אינן מבחינות בין התרומות השונות של התאים השונים בטבלה כך שלמשל, איבר אינדקס יכול להיות מסנן טוב היות והוא מופיע באופן שוטף במסמכים לא רלוונטיים וכמעת ואינו מופיע במסמכים רלוונטיים. ולכן שקלול איבר אינדקס באופן פרופורציונאלי לכוח הסינון שלו בכל עת שהוא מופיע במסמך הוא הדבר השגוי לעשותו במקרה זה. כלומר יש להשתמש במידע שבטבלת השכיחויות כשיוצרים תוכנית שקלול.

היפותזת תוספת הסינון

בפסקה הקודמת, הנחתי הנחה על עצמאות או תלות בצורה פשוטה. או שהנחתי עצמאות של W1 ו W2 או תלות של שניהם. אך כפי שנאמר קודם, זאת אינה הדרך היחידה של הנחת ההנחות הללו. רוברטסון וספארק ג'ונס טוענים שהנחת תלות או עצמאות במסמכים כלשהם יכולה להוביל למסקנה בנוגע לאוסף המסמכים כולו. כדי לראות זאת, ניקח שני אינדקסים i, j כאשר :

$P(xi, xj) = P(xi, xj /w1)P(w1) + P(xi, xi /w2)\ P\ (w2)$

$P(xi)\ P(\ xj) = [P(xi\ /w1)P(w1) + P(xi,\ w2)\ P\ (w2)]\ [P(xj\ /w1)\ P(w1) + P(xj,w2)\ P\ (w2)]$

לעצמאות בלתי מותנית, צריך שיתקיים :

$P(xi, xj) = P(xi, /w1)\ P(xj, w1)\ P(w1) + P(xi\ /w2)\ P(xj/\ w2)\ P\ (w2)$

וזה קורה רק כאשר p(w1)=0 או כאשר p(xi|w1)=p(xi|w2) או כאשר p(xj|w1)=p(xj|w2) . ואם ננסח זאת במילים, כאשר לפחות איבר אינדקס אחד הנו חסר תועלת בסינון בין מסמכים רלוונטיים ולא רלוונטיים. ובאופן כללי עצמאות מותנית תוביל לתלות בלתי מותנית. כעת, נניח שאיברי האינדקס הם אכן עצמאים בצורה מותנית, נקבל את התוצאות יוצאות הדופן הבאות :

קנדל וסטוארט מגדירים מקדם מתאם חלקי בין שני ערכים כך :

$$\rho(X, Y/W) = \frac{\rho(X, Y) - \rho(X, W)\,\rho(Y, W)}{(1 - \rho(X, W)^2)^{1/2}\,(1 - \rho(Y, W)^2)^{1/2}}$$

כאשר [[rho]](.,.|w) ו [[rho]](.,.) הנם בהתאמה מקדמי המתאם. כעת אם X ו Y הנם עצמאים בצורה מותנית, אז :

[[rho]] $(X, Y/W) = 0$

שמוכיח, כאשר משתמשים בביטוי על המתאם החלקי, ש :

[[rho]] $(X, Y) = [[rho]]\ (X, W)\ [[rho]]\ (Y, W)$

היות ו :

$|\ [[rho]]\ (X, Y)\ | <= 1\ ,\ |\ [[rho]]\ (X, W)\ | <= 1\ ,\ |\ [[rho]]\ (Y, W)\ | <= 1$

זה כשלעצמו, מוכיח שכשמשתמשים בהיפותזת העצמאות המותנית :

$|\ [[rho]]\ (X, Y)\ | < |\ [[rho]]\ (X, W)\ |$ or $|\ [[rho]]\ (Y, W)\ |$ (**)

ולכן, אם W הנו משתנה מקרי המייצג את הרלוונטיות, אזי המתאם בינו לבין כל אחד מאיברי האינדקסים גדול מהמתאם בין איברי האינדקסים.

באופן איכותי, אנסה להכליל זאת לפונקציות נוספות בנוסף למדדי מתאם. לינפוט, מגדיר סוג של אמצעי של מתאם התורם מידע :

$rij = (1 - \exp(-2I\ (xi, xj)\ ))^{[1/2\ ]}\ 0 <= rij <= 1$

או :

$$I(x_i, x_j) = -\log\frac{(1 - r_{ij}^2)}{2}$$

כאשר I(xi,xj) הנו האמצעי המוכר למדידת המידע ההדדי. אבל rij מוריד את מקדם המתאם הסטנדרטי [[rho]](.,.) אם (xi,xj) מתפלג נורמאלית. לכן זה הגיוני להניח שעבור התפלגות שאינה נורמאלית rij יתנהג בצורה דומה ל[[rho]](.,.) ויהיה גם הוא מספק. אבל rij מונוטונית עבור l(xi,xj) ולכן גם היא מספקת. לכן אנו יכולים לומר כעת שתחת עצמאות מותנית, המידע

הגלום באיבר אינדקס אחד על איבר אינדקס אחר הוא פחות מהמידע שיש בכל אחד משני האיברים לגבי המשתנה המותנה W. או בניסוח מתמטי:

$I(xi, xj) < I(xi, W)$ or $I(xj, W)$,

כאשר i(.,w) הוא רדיוס המידע, עם המשקלים שלו המבוטאים כהסתברויות פריאוריות. ונזכור כי l(.,w) הוצע כמדד לכוח הסינון. אני סבור שיש לבטא מסקנה זאת כהיפוטיזה, כאשר W מבוטה כתלות.

היפותזת תוספת הסינון: תחת ההיפוטיזה של עצמאות מותנית, המידע הסטטיסטי הגלום באיבר אינדקס אחד על איבר אינדקס אחר פחות מהמידע המגולם בכל אחד מהם על הרלוונטיות.

עלי להדגיש שהטענה שהובילה להיפוטיזה אינה מוכחת. הטענה היא טענה איכותית בלבד. מלבד זאת, הטענה מספקת, בתוספת להיפוטיזה, צידוק מסוים ובסיס תיאורטי לשימוש ב MST המבוסס על I(xi,xj) לשיפור אחזור המידע. היפוטיזת הסינון היא דרך לחזק את היפוטיזת האסוציאציה תחת עצמאות מותנית.

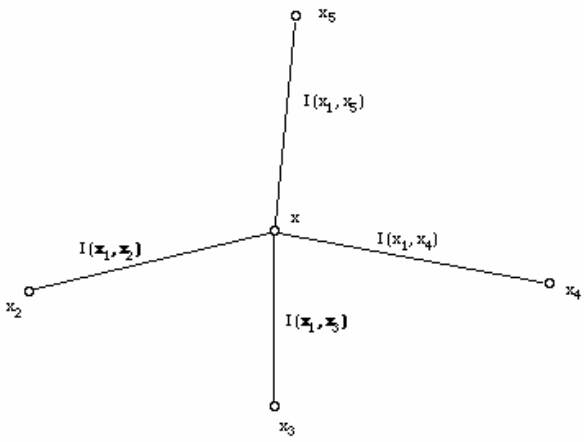

Figure 6.2

תוצאה אחת של היפותזת הסינון, היא שהיא מספקת בסיס הגיוני לדירוג איברי אינדקס הקשורים לשאילתה, בעץ תלות בסדר של I(איבר אינדקס, איבר שאילתה) ערכים שישקפו את סדר ערכי כוח הסינון. הבסיס של זה הוא, שככל שאיבר אינדקס קשור לאיבר שאילתה (שנמדד ע"י EMIM) כך יש סיכוי שכוח הסינון שלו יהיה גדול יותר. כדי להבהיר את הנקודה, נשתמש בדוגמה שנידונה בתרשים 6.2 שלמעלה. נניח ש X1 מייצג איבר שאילתה וש:

$I(x1, x2) < I(x1, x3) < I(x1, x4) < I(x1, x5)$

אז, אומרת ההיפוטיזה שלנו, שמבלי לדעת מראש מה כוח הסינון של כל אחד מאיברי האינדקס 2,3,4,5 זה הגיוני להניח ש:

$I(x2, W) < I(x3, W) < I(x4, W) < I(x5, W)$.

ברור שאיננו יכולים להבטיח סדר זה, אך זהו הניחוש הטוב ביותר בהתחשב בבורות שלנו.

הערות ביבליוגרפיות

הרקע התיאורטי של פרק זה נמצא רק בכמה מאמרים. האחד עוסק בגישה של נתינת משקל הסתברותית המבוססת על מידע על רלוונטיות שנגזר מעבודתו של יו ושותפיו. האחר הוא מאמרם המפורסם של רוברטסון וספארק ג'ונס. לרוע המזל שני המאמרים הללו מבוססים על ההנחה של עצמאות סטוכסטית. המאמר שלי ומאמרם של בוקשטיין וקראפט הם שני המאמרים היחידים שידועים לי שמנסים לבנות מודל בלי הנחה זאת. יתכן שיש לציין גם את מאמרו המוקדם של נגוטיה שמנסה לנקוט בגישה לא ליניארית. סיכומו של רוברטסון נותן מידע על כמה התפתחויות חדשות בנושא.

לפי דויל, היו מרון וקונס הראשונים שתארו שימוש באסוציאציות של איברי האינדקס לצורך ייעול החיפוש. עם זאת, דויל עצמו עבד בשנות החמישים על רעיונות דומים, וכתב כמה מאמרים בנושא. בשנת 1961 סיפק סטילס, שכבר היה מודע לעבודתם של מרון וקונס, פרוצדורה של שימוש בחזרה כפולה בחיפוש, בצורה דומה לשיטה המבוססת על עץ התלות. הוא גם השתמש ב[[chi]] כדי למדוד אסוציאציה בין איברי אינדקס , שמבחינה מתמטית דומה לשימוש במידת חזרת המידע המשותף הנצפית. אם זאת, יש להעדיף את השיטה האחרונה, כשמודדים תלות (ראה גודמן וקרוסקל). סטילס עצמו היה מאוד נחרץ לגבי התועלת שבשימוש באסוציאציות בין איברי אינדקס. הוא ראה שבאמצעותם, ניתן לאתר מסמך הרלוונטי לבקשה גם אם למסמך עצמו לא היה אינדקס זהה לזה שבבקשה.

המודל בפרק זה קשור לעוד שני רעיונות שנידונו במחקרים מוקדמים. האחד הוא משקל הצפיפות של מסמכים הפוכים שכבר נידון בפרק 2 והשני הוא קיבוץ איברים. אם נתייחס תחילה לרעיון הראשון, שמגיע עד למחקרם המוקדם של אדמונסון וויליס, ניתן לכתוב:

$$P(\text{relevance}/\text{document}) \propto \frac{1}{P(\text{document})}$$

במילים: עבור כל מסמך, ההסתברות לרלוונטיות קשורה ביחס הפוך להסתברות של חזרה אקראית. אם מניחים ש p(document) הוא המכפלה של איברי אינדקס בודדים שקיימים או חסרים במסמך, הרי שלאחר חישובים מתאימים מקבלים את כלל המשקל של צפיפות המסמך. הוא מניח את הסבירות הגבוהה ש p(document relevance) קבוע עבור כל המסמכים. השאלה מדוע עקרון זה עובד בצורה כה טובה אינה ברורה אך יתכן שיש לה תשובה בעבודה האחרונה של יו וסלטון.

הקשר עם קיבוץ האיברים נעשה עוד בתחילת הפרק. ניתן להתייחס לעץ המפתח בתור סיווג לאיברי אינדקס. אחת ההשלכות החשובות של המודל המתואר בפרק זה היא שיש פירות ברור כיצד להשתמש בעץ באחזור מידע. עבודות קודמות בתחום התייחסו למקרים פרטיים ולא הובילו למסקנות ממצות.

זה כבר זריך להיות ברור, שהמודל הכמותי מכיל בתוכו מספר נושאים כגון קיבוץ איברים, ניתוח אסוציאציות, משקל הצפיפויות במסמך ומשקל רלוונטיות.

References

1. ROBERTSON, S.E. and SPARCK JONES, K., 'Relevance weighting of search terms', *Journal of the American Society for Information Science*, **27**, 129-146 (1976)

2. van RIJSBERGEN, C.J., 'A theoretical basis for the use of co-occurrence data in information retrieval', *Journal of Documentation*, **33**, 106-119 (1977).

3. BOOKSTEIN, A. and KRAFT, D., 'Operations research applied to document indexing and retrieval decisions', *Journal of the ACM*, **24**, 410-427 (1977).

4. MARON, M.E., 'Mechanized documentation: The logic behind a probabilistic interpretation', In: *Statistical Association Methods for Mechanized Documentation* (Edited by Stevens *et al.*) National Bureau of Standards, Washington, 9-13 (1965).

5. OSBORNE, M.L., 'A Modification of Veto Logic for a Committee of Threshold Logic Units and the Use of 2-class Classifiers for Function Estimation', Ph.D. Thesis, Oregon State University (1975).

6. GOOD, I.J., *Probability and the Weighting of Evidence*, Charles Griffin and Co.Ltd., London (1950).

7. ROBERTSON, S.E., 'The probability ranking principle in IR', *Journal of Documentation*, **33**, 294-304 (1977).

8. GOFFMAN, W., 'A searching procedure for information retrieval', *Information Storage and Retrieval*, **2**, 294-304 (1977).

9. WILLIAMS, J.H., 'Results of classifying documents with multiple discriminant functions', In : *Statistical Association Methods for Mechanized Documentation* (Edited by Stevens et al.) National Bureau of Standards, Washington, 217-224 (1965).

10. DE FINETTI, B., *Theory of Probability*, **Vol. 1**, 146-161, Wiley, London (1974).

11. KU, H.H. and KULLBACK, S., 'Approximating discrete probability distributions', *IEEE Transactions on Information Theory,* **IT-15**, 444-447 (1969).

12. KULLBACK, S., *Information Theory and Statistics*, Dover, New York (1968).

13. CHOW, C.K. and LIU, C.N., 'Approximating discrete probability distributions with dependence trees', *IEEE Transactions on Information Theory,* **IT-14**, 462-467 (1968).

14. COX, D.R., 'The analysis of multivariate binary data', *Applied Statistics*, **21**, 113-120 (1972).

15. BOX, G.E.P. and TIAO, G.C., *Bayesian Inference in Statistical Analysis*, 304-315, Addison-Wesley, Reading, Mass. (1973).

16. GOOD, I.J., *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, The M.I.T. Press, Cambridge, Mass. (1965).

17. HARPER, D. and van RIJSBERGEN, C.J., 'An evaluation of feedback in document retrieval using co-occurrence data', *Journal of Documentation*, **34**, 189-216 (1978).

18. HUGHES, G.F., 'On the mean accuracy of statistical pattern recognizers', *IEEE Transactions on Information Theory*, **IT-14**, 55-63 (1968).

19. MARON, M.E. and KUHNS, J.L., 'On relevance, probabilistic indexing and information retrieval', *Journal of the ACM*, **7**, 216-244 (1960).

20. IVIE, E.L., 'Search Procedures Based on Measures of Relatedness Between Documents', Ph.D. Thesis, M.I.T., Report MAC-TR-29 (1966).

21. WHITNEY, V.K.M., 'Minimal spanning tree, Algorithm 422', *Communications of the ACM*, **15**, 273-274 (1972).

22. BENTLEY, J.L. and FRIEDMAN, J.H., *Fast Algorithm for Constructing Minimal Spanning Trees in Coordinate Spaces*, Stanford Report, STAN-CS-75-529 (1975).

23. CROFT, W.B., 'Clustering large files of documents using single link', *Journal of the American Society for Information Science*, **28**, 341-344 (1977).

24. SALTON, G., *Dynamic Information and Library Processing,* Prentice-Hall, Englewoods Cliffs, NJ., 441-445 (1975).

25. JARDINE, N. and SIBSON, R., *Mathematical Taxonomy,* pp. 12-15, Wiley, London and New York (1971).

26. KENDALL, M.G. and STUART, A., *Advanced Theory of Statistics,* **Vol. 2**, 2nd ed., Griffin, London (1967).

27. LINFOOT, E.H., 'An informational measure of correlation', *Information and Control*, **1**, 85-89 (1957).

28. YU, C.T. and SALTON, G., "Precision Weighting - An effective automatic indexing method', *Journal of the ACM*, **23**, 76-85 (1976).

29. YU, C.T., LUK, W.S. and CHEUNG, T.Y., 'A statistical model for relevance feedback in information retrieval', *Journal of the ACM*, **23**, 273-286 (1976).

30. NEGOITA, C.V., 'On the decision process in information retrieval', *Studii si cercetari de documentare*, **15**, 269-281 (1973).

31. ROBERTSON, S.E., 'Theories and models in information retrieval', *Journal of Documentation*, **33**, 126-148 (1977).

32. DOYLE, L.B., *Information Retrieval and Processing*, Melville Publishing Co., Los Angeles, California (1975).

33. DOYLE, L.B., 'Programmed interpretation of text as a basis for information retrieval systems', In: *Proceedings of the Western Joint Computer Conference*, San Francisco, 60-63 (1959).

34. DOYLE, L.B., 'Semantic road maps for literature searchers', *Journal of the ACM,* **8**, 553-578 (1961).

35. DOYLE, L.B., 'Some compromises between word grouping and document grouping', In: *Statistical Association Methods for Mechanized Documentation* (Edited by Stevens *et al*.) National Bureau of Standards, Washington, 15-24 (1965).

36. STILES, H.F., 'The association factor in information retrieval', *Journal of the ACM*, **8**, 271-279 (1961).

37. GOODMAN, L. and KRUSKAL, W., 'Measures of association for cross-classifications', *Journal of the American Statistical Association*, **49**, 732-764 (1954).

38. EDMUNDSON, H.P. and WYLLYS, R.E., 'Automatic abstracting and indexing - Survey and recommendations', *Communications of the ACM*, **4**, 226-234 (1961).

39. YU, C.T. and SALTON, G., 'Effective information retrieval using term accuracy', *Communications of the ACM*, **20**, 135-142 (1977).

40. SPARCK JONES, K., *Automatic Keyword Classification for Information Retrieval*, Butterworths, London (1971).