PENGEMBANGAN PENDEKATAN MATRIKS HUBUNGAN UNTUK PENGUKURAN SIMILARITAS

Fera Anugreni[1], Herman Mawengkang[2], Marwan Ramli[3]
Programstudi Pascasarjana Teknik Informatika Universitas Sumatra Utara[1]
E-mail: anugreni@yahoo.com
Dosen Dept.Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam USU[2]
Jl. Bioteknologi No.1 Kampus USU
Dosen Dept.Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam USU[3]
Jl. Bioteknologi No.1 Kampus USU

Abstract

The similarity calculation is a basic function which is used by a wide range of information processing applications such as searching for clustering and classification. There are different types of relationships can be used to reinforce the value of the similarity of data objects in pairs and to help improve the quality of the calculation of similarity between data objects. It is helpful to keep track of how relationships between data objects used to measure the similarity of data objects. Results of this study can showed that the Unified Relationship Matrix (URM) can be used to provide heterogeneous data objects and the linkages between each object data in an integrated manner.

Keyword : Similarity, URM, SimFusion

Abstrak

Perhitungan similaritas adalah fungsi dasar yang digunakan oleh berbagai aplikasi pengolahan informasi seperti mencari *clustering* dan klasifikasi. Ada berbagai jenis hubungan dapat digunakan untuk memperkuat nilai kesamaan objek data berpasangan dan untuk membantu meningkatkan kualitas perhitungan kesamaan antara objek data. Hal ini membantu untuk melacak bagaimana hubungan antara objek data yang digunakan untuk mengukur kesamaan objek data. Hasil penelitian ini dapat menunjukkan bahwa Unified Relationship Matrix (URM) dapat digunakan untuk menyediakan data objek heterogen dan hubungan antara masing-masing objek data secara terpadu.

Kata Kunci : Similaritas, URM, SimFusion

1. Pendahuluan

Dalam perkembangan pesat teknologi informasi pada dekade ini, masyarakat terpapar terhadap volume informasi yang semakin hari bertambah besar. Tentunya timbul pertanyaan bagaimana pemakai dapat secara efektif memanfaatkan dan mengintegrasikan volume yang semakin besar ini.

Dokumen, pengguna, metadata, dan jenis-jenis entitas lain ditemukan dalam domain ilmiah/akademis, yang dapat ditemukan di web domain, semua ini dianggap sebagai objek data yang berisi informasi. Informasi demikian ini dapat mencirikan fitur konten individu benda, serta hubungan antara obyek, dari sama atau berbagai jenis sumber.

Banyak karya telah berfokus pada menggunakan satu jenis hubungan ketika menghitung kesamaan objek data. Pendekatan yang menghitung kesamaan obyek dokumen-query menggunakan hubungan jangka dokumen meliputi: Model Ruang Vektor (VSM) [1], Generalized Model Ruang Vektor [2], Semantic Indexing [3], ekspansi permintaan, dan ruang vektor dinamis modifikasi[4].

Tujuan penelian untuk mendapatkan nilai kemiripan dokumen yang lebih terfokus melalui kueri yang diberikan.

2. Metode dan Peralatan

2.1. URM

Definisi formal Unified Relationship Matrix (URM) yang mewakili kedua hubungan antar dan intra-jenis antara objek data heterogen secara terpadu diberikan di bawah ini. Misalkan ada t ruang data yang berbeda S_1, S_2 ... S_t data dalam ruang data yang sama yang terhubung melalui intra-jenis hubungan

$R_i \subseteq S_i x S_i$. Objek data dari dua ruang data yang berbeda terhubung melalui hubungan antar-jenis $R_{ij} \subseteq S_i x S_j \ (i \neq j)$. itu intra-jenis hubungan R_i dapat direpresentasikan sebagai *mxm* matriks ketetanggaan L_i (*m* adalah jumlah objek dalam ruang data S_i). Di dalam matriks L_i sel l_{xy} mewakili hubungan antar-jenis dari obyek x_{th} ke objek y_{th} di ruang data S_i. intertype The Hubungan R_{ij} dapat direpresentasikan sebagai adjacency mxn matriks L_{ij} (m adalah jumlah objek dalam S_i, dan n adalah total jumlah objek dalam S_j), dimana nilai l_{xy} sel merupakan hubungan antar-jenis dari objek X_{th} di S_i ke objek j di S_j. Jika kita menggabungkan N ruang data menjadi terpadu ruang data U, maka hubungan antar dan intra-tipe sebelumnya adalah bagian dari intratype hubungan Ru di ruang data U. Misalkan L_u adalah adjacency matriks R_u, maka L_u adalah matriks persegi. mendefinisikan Unified. L_{urm} sebagai matriks yang menggabungkan semua matriks hubungan, seperti yang diberikan dalam persamaan.

$$L_{urm} = \begin{vmatrix} L_{11} & L_{12} & \cdots & L_{1N} \\ L_{21} & L_{22} & \cdots & L_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ L_{N1} & L_{N2} & \cdots & L_{NN} \end{vmatrix} \quad (1)$$

The ***URM*** menyediakan cara umum memandang objek data dan hubungan mereka. Di ***URM*** itu, berbagai jenis objek yang diperlakukan sebagai elemen dari ruang "bersatu" data. Antar- sebelumnya dan hubungan intra-jenis masing-masing dianggap sebagai intra-jenis generik hubungan yang menghubungkan objek data dalam ruang "bersatu" data. ***URM*** yang dapat digunakan untuk menjelaskan berbagai dunia nyata skenario aplikasi informasi. Sebagai contoh, jika hanya mempertimbangkan satu ruang data, halaman web, dan satu jenis intra-jenis hubungan, hubungan hyperlink, maka ***URM*** berkurang untuk matriks hubungan ketetanggaan dari graf web. Pertimbangkan contoh lain. Jika memiliki dua ruang data, dokumen dan istilah, maka hubungan antar-jenis didefinisikan ketika Dokumen berisi istilah atau istilah yang terkandung oleh dokumen. SEBUAH ***URM*** dapat dibangun seperti pada persamaan:

$$L_{urm} = \begin{vmatrix} 0 & L_{dt} \\ L_{dt}^T & 0 \end{vmatrix} \quad (2)$$

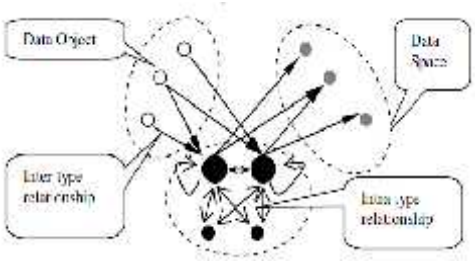
L_{dt} adalah matriks jangka dokumen tradisional yang digunakan dalam VSM. 0 sub-matriks pada diagonal menunjukkan bahwa tidak memiliki sebelumnya pengetahuan tentang hubungan intra-jenis dalam dokumen atau jangka ruang. Semua aplikasi informasi yang memanipulasi matriks jangka dokumen masih dapat digunakan pada L_{urm}. Selain itu, hubungan intra-jenis ruang dokumen dan jangka bisa diperoleh hanya dengan mengalikan L_{urm} dengan dirinya sendiri:

$$L'_{urm} = L_{urm} \times L_{urm} = \begin{vmatrix} 0 & L_{dt} \\ L_{dt}^T & 0 \end{vmatrix} \quad (3)$$

Di mana *Ld* dan *L t* sesuai dengan dokumen pair-bijaksana matriks kesamaan dan jangka berpasangan kesamaan matriks diperoleh oleh sebagian besar perhitungan kesamaan VSM tradisional. Oleh menambahkan L'_{urm} dan L_{urm}, dapat memiliki ***URM*** lengkap untuk dokumen dan jangka spasi: $\begin{vmatrix} Ld & L_{dt} \\ L_{dt}^t & L_t \end{vmatrix}$ Matriks tertentu yang menggabungkan hubungan dokumen pair-wise dan jangka berpasangan dengan hubungan jangka dokumen tradisional disebut oleh Davidson sebagai "generik matriks yang diperbesar"

2.2. Similarity

Asumsi dasar adalah bahwa: "kesamaan antara kedua data benda dapat diperkuat oleh kesamaan objek data terkait dari ruang yang sama dan berbeda ", seperti yang digambarkan di bawah ini:



Gambar 1. ilustrasi asumsi penguatan similaritas[5]

Seperti dapat dilihat pada Gambar 1, kesamaan antara kedua data benda (besar simpul hitam di tengah) diperkuat oleh hubungan kapal dari jenis yang sama dari objek data terkait (kecil hitam node) serta hubungan (baik inbound dan outbound) dari berbagai jenis objek data (putih dan bening abu-abu).

Misalkan ada N ruang data yang berbeda X_1, X_2, ... X_N. Data benda-benda di ruang yang sama terkait melalui hubungan intra-jenis $R_i \subseteq X_i x X_i$. Objek data dari ruang yang berbeda terkait melalui intertype hubungan $R_{ij} \subseteq X_{ij} x X_{ij} \ (i \neq j)$. Hubungan yang dianggap serupa di alam. S_{ij} *(x, y)* merupakan kesamaan antara objek *x* dari ruang *i* dan *y* objek dari ruang *j*. R_{ij} *(x, y)* mewakili antar-*(i = j)* atau intra-jenis $(i \neq j)$ hubungan dari objek *x* dalam ruang *i* keberatan *y* dalam ruang *i*, sedangkan *a* dan *b* adalah data benda dalam ruang data di bawah kondisi bahwa *x* adalah terkait dengan *a* dan *y* berhubungan dengan *b*. Kemudian asumsi penguatan kesamaan matematis dapat disajikan sebagai:

(3.2)

$$S_{ij}^{new}(x,y) = S_{ij}^{original}(x,y) + \sum_{(\forall k, and\ k \neq i),(\forall l, and\ l \neq j)} R_{ik}(x,a) R_{jl}(y,b) S_{kl}^{original}(a,b) \quad (4)$$

di mana *a* dan *b* adalah parameter positif digunakan untuk mengatur (selama Proses penguatan) kepentingan relatif yang asli kesamaan obyek *x* dan *y* dengan pentingnya kesamaan diperkuat oleh hubungan antar dan intra-jenis. Jika menggunakan satu set parameter positif $^{a}_{ij}$ untuk mewakili kepentingan relatif dari kesamaan diperkuat dari data ruang *i* ke ruang data *j*, dan mempertimbangkan jumlah nilai kemiripan asli yang terlibat dalam Proses sebagai nilai kesamaan diperkuat melalui intra-tipe khusus hubungan yang mengarah ke data obyek itu sendiri, asumsi penguatan kesamaan dapat direpresentasikan sebagai:

$$S_{ij}^{new}(x,y) = {}^{a}_{ii} R_{ii}(x,x)\, {}^{a}_{jj} R_{jj}(x,x) S_{ij}^{original}(y,y) + \sum_{(\forall k, and\ k \neq i),(\forall l, and\ l \neq j)} R_{ik}(x,a) R_{jl}(y,b) S_{kl}^{original}(a,b) \quad (5)$$

$$S_{ij}^{new}(x,y) = \sum_{\forall a, \forall b} {}^{a}_{ik} R_{ik}(x,a)\, {}^{a}_{jl} R_{jl}(y,b) S_{kl}^{original}(a,b) \quad (6)$$

Mengingat obyek terkait satu objek data dalam ruang data lainnya sebagai pemetaan di ruang-ruang data, alasan kesamaan Proses penguatan dapat menghasilkan perkiraan yang lebih baik adalah bahwa kesamaan dua objek data diukur dalam berbagai perspektif (tempat data) bukan perspektif tunggal. Namun, pra- syarat adalah bahwa hubungan yang terlibat akurat dan aditif. Dengan demikian, perawatan harus dilakukan untuk menghindari melibatkan bertentangan atau jenis ambigu hubungan

2.3 Sim Fusion

Misalkan ada ruang yang berbeda N sedang dipertimbangkan, dan URM dikembangkan alam cara yang mirip dengan untuk mewakili hubungan antar dan intra-jenis seperti yang ditunjukkan dalam Persamaan dibawah ini:

$$L_{urm} = \begin{vmatrix} {}^{a}_{11} L_{11} & {}^{a}_{12} L_{12} & \cdots & {}^{a}_{1N} L_{1N} \\ {}^{a}_{21} L_{21} & {}^{a}_{22} L_{22} & \cdots & {}^{a}_{2N} L_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ {}^{a}_{N1} L_{N1} & {}^{a}_{N2} L_{N2} & \cdots & {}^{a}_{NN} L_{NN} \end{vmatrix} \quad (7)$$

Dimana Li adalah hubungan matriks intra-jenis data ruang *i* dan L_{ij} adalah hubungan antar matriks-jenis dari ruang data *i* ke ruang data j. Jumlah setiap baris dari salah satu sub-matriks adalah dinormalkan ke 1. Dalam kasus yang objek data *x* dari ruang tidak memiliki hubungan dengan setiap objek data di ruang data *j* (semua elemen dalam baris ke-L_{ij} dari matriks adalah nol), maka setiap elemen dalam engan baris hubungan matriks L_{ij} diatur ke *1/n*, dimana *n* adalah total jumlah elemen dalam ruang *j*. Ini sama dengan menggunakan Hubungan acak untuk mewakili no-hubungan. Kami juga mendefinisikan set parameter ^{a}S untuk menyesuaikan kepentingan relatif dari berbagai hubungan antar dan intra-jenis, sehingga untuk setiap i, $\sum_{\forall j} {}^{a}_{ij} = 1$ dan $\forall i, j\ {}^{a}_{ij} > 0$. Jadi, Persamaan adalah deretan-stokastik matriks dan dapat diberikan sebagai probabilitas langkah tunggal matriks transformasi dalam Rantai Markov. Kami juga mendefinisikan Unified Similarity Matriks (USM), Susm, untuk mewakili nilai-nilai kesamaan dari setiap pasangan objek data dari sama atau ruang data yang berbeda pada awal algoritma:

$$S_{usm}^{new} = L_{urm} S_{usm}^{original} L_{urm}^{T} \quad (8)$$

$$S_{usm}^{n} = L_{urm} S_{usm}^{n-1} L_{urm}^{T} = L_{usm}^{n} S_{usm}^{0} (L_{urm}^{T})^{n} \quad (9)$$

Seringkali tindakan kesamaan didefinisikan sebagai penurunan fungsi tions dari metrik jarak. Sebagai contoh, dua string metrik paling sering digunakan adalah editDistance (Lin 1998) dan tri gram (Bahl, Jelinek, dan Mercer 1983). Untuk string yang terbatas x dan y jarak mengedit didefinisikan sebagai

$$d_{edit}(x,y) = \min\{¥(S) | S \text{ is a en edit sequence taking x to y}\} \quad (10)$$

dimana γ menunjukkan biaya operasi mengedit (penghapusan, penyisipan, penggantian), dan untuk urutan mengedit opera- otions $S = \{s_1, \ldots, s_n\}$, $¥(S) = \sum_{i=1}^{n} ¥(s_i)$. Trigram yang jarak untuk dua urutan x dan y didefinisikan sebagai:

$$d_{tri}(x,y) = \frac{|tri(x) \cap tri(y)|}{|tri(x) \cup tri(y)|} \quad (11)$$

mana tri (x) menunjukkan koleksi trigram (memerintahkan substring panjang 3) dari x, dan | tri (x) | menunjukkan nomor dari trigram x. Kemudian langkah-langkah kesamaan berkoresponden didefinisikan sebagai persamaan :

$$sim_a(x,y) = \frac{1}{1 + d_a(x,y)} \quad (12)$$

Dimana $a \in \{edit, tri\}$

Sejak L_{URM} dapat dianggap sebagai matriks langkah transisi tunggal dari Markov Chain, proses penguatan kesamaan berulang dapat dijelaskan dalam "dua model yang walker acak". Misalkan dua pejalan kaki secara acak mulai dari dua objek data dalam ruang terpadu dan mereka berjalan dari satu objek ke yang lain, langkah demi langkah. Dalam setiap langkah, masing-masing akan memilih objek berikutnya untuk menginjakkan kaki di sesuai dengan distribusi probabilitas bagaimana Data saat ini terkait dengan benda-benda lainnya sebagaimana dimaksud dalam L_{URM}.

Jika *S 0 usm* juga dapat diberikan sebagai objek ke objek hubungan matriks distribusi, maka kesamaan diperkuat antara dua benda asli yang dua pejalan kaki mulai perjalanan mereka, dapat diterjemahkan ke dalam kemungkinan bahwa dua pejalan kaki saling bertemu, setelah masing-masing berjalan *n* langkah sesuai dengan L_{URM} *SimFusion* juga dapat dianggap sebagai grafik spektral generik proses partisi.

3. Hasil dan Pembahasan

Versi yang disederhanakan dari ***SimFusion*** yang hanya mempertimbangkan satu atau dua jenis objek data telah divalidasi melalui studi sebelumnya. Sebagai contoh, perhatikan hanya satu ruang data, dari artikel jurnal, dan salah satu jenis hubungan, hubungan referensi antara artikel jurnal. Mengatur awal S_{usm} menjadi matriks identitas, mengurangi ke co-citation, di mana kesamaan dua artikel adalah ditentukan oleh jumlah artikel mereka berdua mengutip. Jika hanya pertimbangkan hubungan referensi terbalik, mengurangi ke bibliografi kopling, di mana kesamaan dua artikel adalah ditentukan oleh jumlah artikel yang mengutip mereka berdua. mempertimbangkan ***URM*** dalam

$$Lurm=\begin{vmatrix} 0 & L_{dt} \\ L^T{}_{dt} & 0 \end{vmatrix} \tag{13}$$

yang merupakan dokumen ruang, ruang jangka, dan hubungan "yang berisi" dokumen persyaratan. Misalkan tidak memiliki pengetahuan sebelumnya tentang kesamaan setiap objek data dan mengatur S_{usm} menjadi matriks identitas. menerapkan

$$S^n_{usm}=LurmS^{n-1}_{usm}L^T_{urm}=L^n_{urm}S^0_{usm}(L^T_{urm})^n \tag{14}$$

Menghasilkan perhitungan dokumen berpasangan kesamaan dan berpasangan kesamaan jangka sesuai dengan VSM. Itu tetap merupakan masalah yang menarik, apakah memperkaya ***URM*** dan ***USM*** dengan beberapa pengetahuan sebelumnya (misalnya, tesaurus, atau dokumen hubungan referensi) dan iteratif memperkuat kesamaan, akan menghasilkan pengetahuan yang lebih baik dari kesamaan istilah, kesamaan dokumen, dan persamaan jangka dokumen.

Baru-baru ini, telah mencoba untuk mengakses halaman query web page hubungan untuk lebih memprediksi kesamaan objek web sehingga untuk membantu meningkatkan efektivitas algoritma web-pengelompokan. Metode mereka untuk menghitung kesamaan web objek juga dapat dengan baik dimodelkan oleh algoritma ***SimFusion***. Misalkan ada dua ruang Data: ruang halaman Web dan ruang query. Dua ruang dimodelkan dalam ***URM*** sebagai:

$$Lurm=\begin{vmatrix} {}^{a}{}_{11}L_{query} & {}^{a}{}_{12}L_{query-page} \\ {}^{a}{}_{21}L_{page-query} & {}^{a}{}_{22}L_{webpage} \end{vmatrix} \tag{15}$$

di mana L_{query} mengacu pada hubungan kesamaan konten permintaan matriks dan $L_{webpage}$ mengacu pada konten halaman web kesamaan hubungan matriks.

Jika $L_{webpage}$ mengacu pada query dengan yang daftar pencarian hubungan yang sesuai, dan *Lusm* adalah identitas matrix, ${}^{a}{}_{11}={}^{a}{}_{22}=0$ dan ${}^{a}{}_{12}={}^{a}{}_{21}=1$, kemudian menerapkan Persamaan (14) untuk ini ***URM*** akan menghasilkan Raghavan dan Sever kerja, di mana mereka mengukur kesamaan query berdasarkan sesuai daftar dokumen hasil.

Jika $L_{webpage}$ mengacu pada web permintaan web Halaman hubungan klik-melalui, dan *Susm* adalah matriks identitas, dan semua λs tetap sama, menerapkan persamaan (14) ke ***URM*** ini akan mengakibatkan Beeferman dan Berger metode pengelompokan, di mana mengukur kesamaan query menggunakan kesamaan diklik halaman web dan menghitung kesamaan halaman web menggunakan kesamaan permintaan yang menyebabkan pemilihan web halaman. Jika kita ${}^{a}{}_{11}>0, {}^{a}{}_{22}>0, {}^{a}{}_{12}>0$ dan ${}^{a}{}_{21}>0$, menerapkan persamaan (14) untuk ***URM*** ini akan menghasilkan karya, di mana permintaan kesamaan didasarkan pada kedua kesamaan isi permintaan dan hubungan kesamaan dokumen yang dipilih oleh pengguna yang mengajukan permintaan.

4. Kesimpulan

Dari hasil kajian penelitian dapat disimpulkan bahwa Unified Relationship matrix (URM) dapat dipakai untuk menyajikan objek data yang heterogen serta keterkaitan antara masing-masing objek data dalam cara yang terpadu. Penelitiam ini juga membicarakan tentang persoalan intgreasi informasi dengan memperlihatkan bgaimana hubungan berbeda dapat dipakai untuk meningkatkan ukuran kesamaan (similaritas) objek data.

Algoritma SimFusion dapat secara efektif memadukan hubungan dari berbagai sumber untuk menunjukkan kesamaan objek data. Algoritma ini dikembangkan berdasarkan URM dari objek data.

Daftar Pustaka

[1] Salton, G. 1968. Automatic Information Organization and Retrieval. McGraw-Hill.

[2] Wong, S.K.M., Ziarko W., Raghavan V.V., & Wong, P. C. N. 1987. .On Modeling of information Retrieval Conscept in Vektor Space. ACM TODS, 12:299-321.

[3] Dumais, S.T., Furnas, G.W., Landaur, T.K., Deerwester, S., Harshman, R. 1988. Using latent semantic analysis to improve information retrieval. *Proceeding of the conference on Human Factor in Computing Systems*. pp. 281-285, Wanshington D.C.

[4] Brauen, T.L. 1971. Documen Vektor Modification, *in The Smart Retrieval System_experiments in Automatic Document Processing*, G. Salton, editor, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, Chapter 24.

[5] W. Xi, B.Zhang & E. A.Fox. (2005). *SimFusion: Measuring Similarity using Unified Relationship Matrix*. SIGIR'05, August 15-19. Salvador, Brazil.

[6] P.D Turner & P.Pantel. (2010). *From Frequency to meaning: vector space models of semantic*. journal of Artificial Research, 37 pp.141-1888,.

[7] Han, J. & Kamber, M. 2006. *Data Mining Concepts and Techniques*. *Second Edition*. Elsevier: USA.

[8] Turban, E., Aronson, J.E. & Liang, T.P. 2005. *Decision Support and Business Intelligence Systems. Seventh Edition*. Pearson Higher Education. USA.

[9] W. Xi, B.Zhang & E. A.Fox. (2004). *SimFusion: A Unifield Similarity Measurement Algorithm for Multy-type Interrelated Web Objects.* Technical Report, TR-04-19, Computer Science Department, Virginia Tech

[10] Kantardzic, M. 2003. *DATA MINING Concepts, Models, Methods, and Algorithms*. IEEE Press: USA.

[11] Ribeiro-Neto, R & Muntz, R. 1996. A Belief Network Model for IR. *Proceeding of the 19th ACM-SIGIR conference on research and development in information retrieval*, pp. 253-260.

.