الطريقة الاحتمالية في اعداد الكشافات اعتمادا على علاقة التشابه الموضوعي مابين الوثائق المشيرة والمشار اليها .

د. نعيمة حسن رزوقي
كلية الاداب / الجامعة المستنصرية

المقدمة

كثيراً مايهمنا في دراسة استرجاع المعلومات علاقة التشابه مابين طالب المستفيد و كل وثيقة بضمن مجموعة معينة من الوثائق وعندما يكون مقياس التشابه عالياً عندئذ نقول بأن الوثيقة ملائمة للطلب وقد تم استرجاعها . وقد يكون مقياس التشابه هذا قائماً على مقدار احتمالية ملاءمة الوثيقة للطلب او على اساس الترتيب التنازلي للوثائق بحسب ملاءمتها حيث تكون الوثائق في الرتبة الاعلى هي الأكثر ملاءمة مقارنة بالرتبة الادنى .

وفي الغالب تعتمد العبارات الدالة او الواصفات كوحدات تتمركز حولها الوثائق المتشابهة للطلبات المتشابهة ، ومن هنا فقد اكد عدد من الباحثين (١ ، ٢) على ان الغرض من دراسة علاقة التشابه مابين الوثائق ضمن مجموعة معينة هو تحقيق نتائج افضل في عملية الأسترجاع ، وقد عبر عن هذه العلاقة بالفرضية العنقودية (Clustering Hypothesis) [3] والتي تنص على انه في الغالب تكون الوثائق المتشابهة ملائمة لنفس الطلبات . وعليه فقد سعت الدراسة الحالية على تطبيق الطريقة الاحتمالية في اختيار العبارات الدالة التي تؤلف الكشاف او الدليل لمجموعة الوثائق وذلك اعتماداً

على علاقة التشابه مابين الوثائق المصدرية (Source Documents) ومجموعة الوثائق المشار اليها . ولتحقيق ذلك فقد اعتمدت مستخلصات المقالات كبديل عن المقالات في اختيار العبارات الدالة التي تمثل محتوى المقالة كما اعتمدت عناوين الوثائق المشار اليها كمصادر للبحث لنفس الغرض وكانت النتيجة لهذا التحليل الحصول على قائمة من العبارات الدالة ومن ثم اجريت المطابقة مابين :

١ — مجموعة العبارات الدالة في عناوين الوثائق المشار اليها في المقالة الواحدة وعبارات مستخلص المقالة بغية التوصل الى عدد العناقيد التي تكونت من تجانس عدد من العناوين ومستخلص المقالة من حيث تشابه العبارات الدالة فيها .

٢ — مجموعة العبارات الدالة في عناوين الوثائق المشار اليها في المقالة الواحدة لغرض التوصل ان عدد العناقيد التي تتكون من تجانس عدد من العناوين فيما بينها من حيث تشابه العبارات الدالة في عناوينها .

ومن ثم احتساب مجموع العناقيد التي بموجبها يتم قياس درجة التشابه مابين الوثائق في المقالة الواحدة التي تمثل طلباً وجواباً للطلب ، ولانجاز هذه الدراسة تم اختيار فرضيتين هما :

١ — ان الوثائق المتشابهة ملائمة على وجه العموم لنفس الطلب .

٢ — ان الوثائق المتشابهة تتجمع بشكل عناقيد حول نفس العبارات الدالة .

هذا وقد اقتصرت الدراسة على احدث الاعداد التي تم الحصول عليها في مجال علم المعلومات وللدورتين الاتيتين :

1. Journal of the American Society for Information Science (JASIS). vol.36 (1985)
2. Journal of Information Science (JIS) vol. 14 (1983) .

حيث بلغ مجموع الاعداد التي تم تحليلها لاغراض الدراسة (10) اعداد ، ستة منها لمجلة (JASIS) والأربعة الأخرى لمجلة (JIS) وكانت الحصيلة الحاصلة من هذه الاعداد العشرة (69) مقالة متضمنة مجموعة من

الوثائق التي تمت الاشارة اليها وحجمها (1597) عنواناً وقد اسفر التحليل لاختيار العبارات الدالة عن مجموع كلي مقداره (367) عبارة دالة من المستخلصات و (2116) عبارة دالة من عناوين الوثائق المشار اليها ، يمثل الجدولان (1) و (2) خلاصة تفصيلية للبيانات المتعلقة بالعينة والتي تدل الرموز فيها على مايأتي :

1 — رق « رقم العدد»

2 — مج م « مجموع المقالات المصدرية في العدد الواحد

3 — مج و « مجموع الوثائق المشار اليها في مجموع مقالات العدد الواحد»

4 — مج مس « مجموع العبارات الدالة في مجموع مستخلصات المقالات في العدد الواحد . »

5 — مج وثق « مجموع العبارات الدالة في مجموع الوثائق المشار اليها في مجموع مقالات العدد الواحد»

6 — مج ع م « مجموع العناقيد المتكونة من ترابط عدد من الوثائق مع المستخلص بعباراتها الدالة»

7 — مج ع و « مجموع العناقيد المتكونة من ترابط عدد من الوثائق في قائمة المراجع بعباراتها الدالة »

8 — مهملة « مجموع عناوين الوثائق المشار اليها والتي اسقطت من التحليل واحتساب التشابه للاسباب الاتية :

أ — تكرارها في مراجع المقال الواحد

ب — عدم دلالة عباراتها

ج — كونها بلغة غير الانكليزية كالفرنسية والالمانية

رق	مج م	مج و	مج مس	مج وثق	مج ع م	مج ع و	مهملة
1	7	96	38	171	23	21	2
2	6	213	34	243	20	46	13
3	9	256	62	311	33	57	18
4	6	170	21	205	11	47	7
5	9	218	47	314	25	54	9
6	9	135	44	171	26	19	16
6	46	1088	246	1415	128	199	65

الجدول رقم (١)

البيانات التفصيلية لمجلة (JASIS)

رق	مج م	مج و	مج مس	مج وثق	مج ع	مج ع و	مهملة
1	7	214	37	284	21	51	5
2	6	161	34	218	1	48	1
3	5	37	23	58	218	3	10
4	5	97	27	141	15	22	/
4	23	509	121	701	55	124	16

الجدول رقم (2)

البيانات التفصيلية لمجلة (JIS)

تعريفات Definition

نظراً لمتطلبات الدراسة في استخدام عدد من المصطلحات وبمعنى محدد توجب اعطاء تعريف مختصر لكل منها على النحو الآتي .

العنقدة (Clustering)

هي عملية تكوين مجموعات متجانسة (Homogenous groups) من الوثائق بالشكل الذي تكون فيها كل وثيقة في المجموعة الواحدة مرتبطة ارتباطاً وثيقاً ببقية الوثائق اقرانها في تلك المجموعة ويقل ارتباطها بالوثائق الاخرى في المجموعات الاخرى ويطلق على كل مجموعة متجانسة بالعنقود [4] كما ان وحدة الارتباط هنا هي العبارة الدالة .

الوثيقة المصدر (Source Document)

هي الوثيقة المتوفرة والمعتمدة في التحليل والتي بدورها قد اشارت الى عدد من الوثائق ، وفي مجال الدراسة الحالية فان الوثيقة المصدر تمثل المقالة في الدورية المتضمنة بالتحليل .

الوثيقة المشار اليها (Cited Document)

هي الوثيقة التي تظهر في قائمة المصادر او المراجع والتي تمت الاشارة اليها من قبل الوثيقة المصدر .

الافتراضات Assumptions

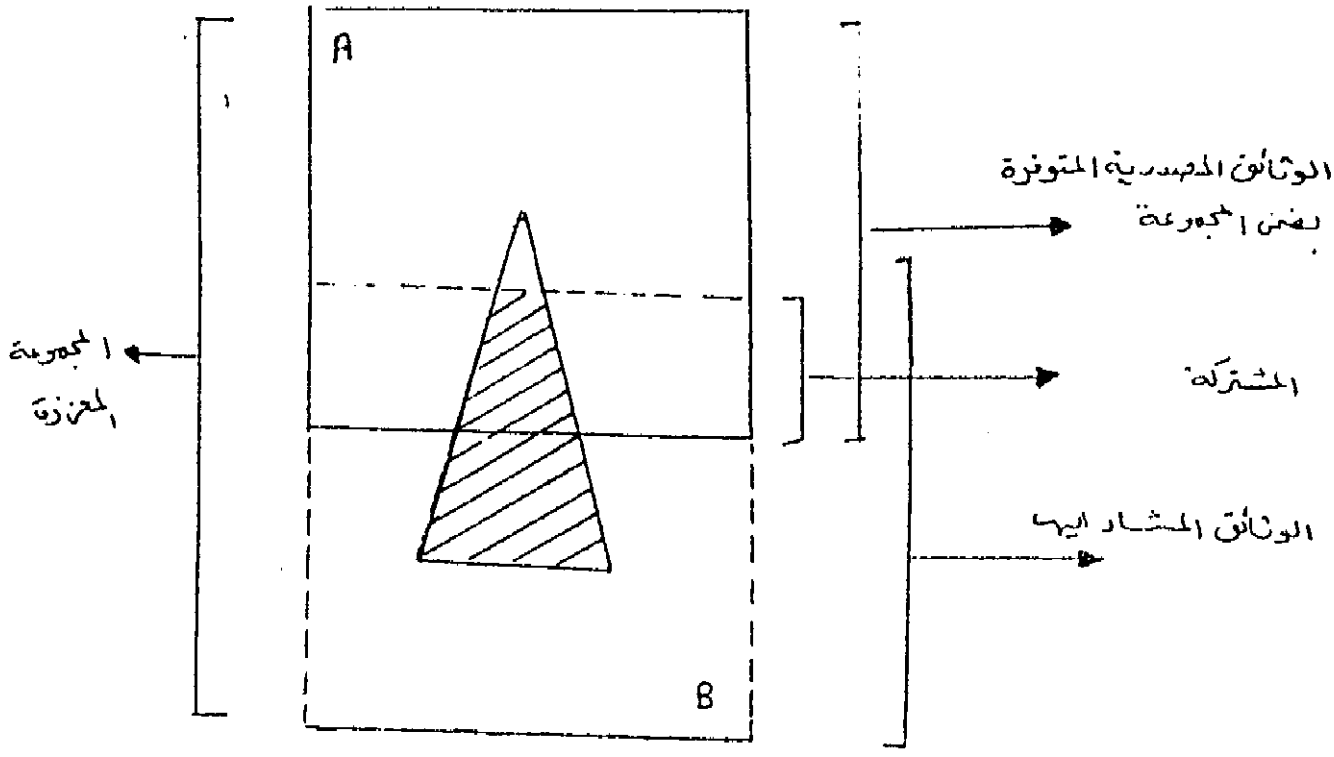
من خلال تفحص ادبيات الموضوع وتحديداً الدراسات السابقة المتعلقة بطرق قياس درجة التشابه بين الوثائق (Documents) والطلبات (queries) تم استخلاص ثلاثة افتراضات ينص الاول منها على ان علاقة التشابه بين الوثائق المشيرة الى (Citing) والمشار اليها (Cited) تشكل مجموعات (عناقيد) من المواد المترابطة والتي يمكن بدورها ان تحقق نظام استرجاع كفء ومدال (7, 6) وعليه فقد دعت تلك الدراسات الى عنقدة الوثائق من اجل تحسين الاسترجاع .

اما الافتراض الثاني فانه يؤكد على ان الوثائق المشار اليها من قبل الوثيقة المصدر تؤلف عينة ملائمة لطلب المؤلف على افتراض سابق مفاده انه غالباً مايشير المؤلف الى الوثائق ذات العلاقة بموضوع الوثيقة المصدر التي اعدها والتي بدورها جميعاً تشكل عنقوداً ملائماً من الوثائق المتشابهة لحاجته من المعلومات ، بعبارة اخرى تمثل اجابات ملائمة لمتطلبات الوثيقة المصدر التي يمكن التعامل معها هنا بمثابة طلب (query) من منطلق ان ارضاء حاجة المستفيد المرتبطة بطلب محدد [7] .

ومن هنا يتبين لنا ان مجموعة الوثائق المتوفرة ممزوجة اساساً بجميع الوثائق التي تمت الاشارة اليها من قبلها فهي بذلك قابلة للزيادة ليس على اساس اضافة مصدر فعلي للمجموعة فحسب بل مع المصدر المضاف قائمة من المصادر الاخرى التي تسند الوثيقة وتعززها وبالتالي تعزز المجموعة ككل والتي يمكن لنا ان نطلق عليها بالمجموعة المعززة او القابلة للزيادة ، ثم هو واضح بالشكل (١) الذي يشير فيه المربع (a) بالخطوط المتقطعة الى مجموع الوثائق التي تمت الاشارة من قبل الوثائق المصدر التي يمثلها المربع (A) اما المثلث في الوسط فأنه يعني مجموع الاجابات لطلب معين . كما يشترك المربعان في جزء منهما ليمثل الوثائق المتوفرة ضمن المجموعة وقد

ظهرت في نفس الوقت ضمن الوثائق المشار اليها . فاذا ماعبرنا عن مجموع الوثائق المصدرية ب (×٤)ومجموع الوثائق المشار اليها لكل وثيقة مصـــدر ب (٤٢) فان المجموعة المعززة (N)تساوي :

$$N = ٤ (× + ٤٢)$$



الشكل رقم (١)

المجموعة المعززة والعلاقة مابين المتوفر والمشار اليه للطلب

اما الافتراض الثالث والاخير فأنه يشير الى ان نجاح مؤلف مافي البحـــث العلمي لها مدلول على ان ذلك المؤلف قد راجع ادبيات الموضوع واعمـــال ذات العلاقات المنجزة حديثاً او قديماً ،وان درجة الشمول لتغطية تلك الأعمال من خلال الاشارة اليها في البحث تعتمد على المؤلف نفسه ، ومع ذلك فمـــن الطبيعي ان تظهر في نهاية البحث او هوامشه قائمة من الوثائق ذات العلاقة

والملائمة التي يراها مناسبة فاشار اليها . وهذا بدوره يعني ان علاقة التشابه قائمة ويمكن ان ينظر اليها على اساس ان بحث المؤلف بمثابة طلب وان الوثائق المشار اليها بمثابة استرجاع وبأي طريقة يختارها المؤلف [8] . وعليه فسان خلف كل وثيقة مصدر (x) هنالك مجموعة من الوثائق الملائمة (٤٢) والتي يكون محتواها الموضوعي هو مايبحث عنه المؤلف ويرغب الكتابة والبحث عليه . هذا وقد اعتبر كووك (kwok) [9] ان مستخلص الوثيقة وعنواها بدائل عن النص الاصلي في تمثيلها للطلبات وبناء على ذلك فسان العبارات الدالة الواصفات التي يتم اختيارها من المستخلص او العنوان يمكن لها ان تمثل تلك الوثائق ليعتمد عليها كروابط تجمع حولها الوثائق المتشابهة لتكون عناقيد وكل عنقود بحد ذاته يؤلف عينة عشوائية لجميع الوثائق الملائمة للطلب بضمن المجموعة المعززة التي اشرنا اليها سابقاً .

نتائج التحليل

بعد تحليل مستخلصات المقالات عينة الدراسة وعناوين الوثائق الملحقة بكل مقالة وتبويب البيانات وجد الاتي :

١ ـ بالرغم من اختلاف في مجموع الاعداد التي تم تحليلها للدورتين وماتبعه من اختلاف في المجموع النهائي للمقالات والوثائق المشار اليها الا ان هناك توافقا في المعدلات النهائية للعبارات والعناقيد في المقال الواحد حيث كان :

ا ـ معدل عدد العبارات الدالة في المستخلص الواحد هو (5) عبارات لكلتا الدورتين .

ب ـ معدل عدد العناقيد المتكونة من مستخلص كل مقالة مصدرية والوثائق التي اشيرت اليها (3) عناقيد في كلتا الدورتين تتمثل بثلاث عبارات دالة .

ج ـ معدل عدد العناقيد المتكونة من الوثائق المشار اليها في المقال الواحد مابين (10) عنقوداً .

وعليه تم الاعتماد في التحليل على المجموع الكلي للبيانات حـــول الدوريتين ولم نجد جدوى من التمييز بينهما .

٢ ـ وجد ان هناك ترابطاً موضوعياً بين قائمة العبارات في المقالة الـواحـدة لتكون عناقيد فرعية صغيرة يمكن ربطها باحالات لتمثل مجموعـــة متجانسة اكبر ذات علاقة بموضوع اساسي . ففي المقالة المنشورة في مجلة علم المعلومات (Jis) [10] هناك ست عبارات دالة تعـدد استخدامها في عناوين الوثائق ومستخلص المقالة وبالتالي فانها مرتبطة موضوعياً بالنظم الصوتية وهذه العبارات هي :

- Speech technolojy
- Voice Systems
- Speech Interface
- Man Machine interface
- Speech recognition
- Human Factor

٣ ـ كما وجد انه كلما زاد عدد الوثائق المشار اليها في المقالة الواحدة تعددت العناقيد و كثرت فيها المترادفات او الصيغ المختلفة للعبارات مثل المفـرد والجمع والمختصرات التي يمكن لها ان ترتبط مع بعضها .

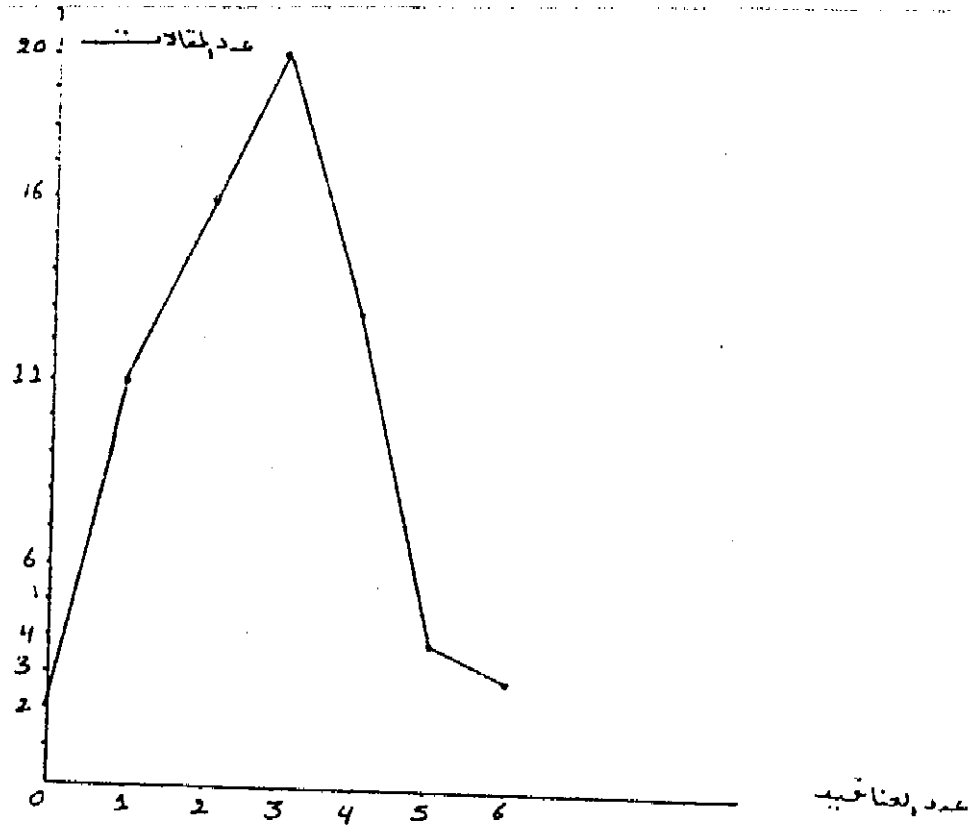
٤ ـ افتقار عدد من المقالات الى الأرتباط مابين عبارات المستخلص والوثائق المشار اليها وقد وجد ان هذا مرتبط بطبيعة المقالة فاذا كانت من المقالات الأستعراضية العامة كان مقدار التشتت في الموضوعات واضحاً لمحاولة موقف المقالة تغطية الموضوع المدروس لكل جوانبه مما يحدث انعـدام الترابط بين العناوين من خلال العبارات الدالة المتشابهة حيث يكـــون عدد العناقيد فيها صفراً او التشتت الواضح لعناوينها وتناثرها في مجموعات صغيرة جداً لاتتعدى ثلاثة عناوين فقط بسبب تعدد العبارات المستخدمة .

ولاثبات ذلك تم احتساب الوسط الحسابي والانحراف المعياري للبيانات الواردة في الجدول (3) الذي يمثل الحقل الاول فيه مجموع المقالات التي يحصل او لايحصل فيها ترابط مابين عبارات مستخلصاتها وعناوين الوثائق المشار اليها في حين يمثل الحقل الثاني منه عدد العناقيد المتنوعة من الترابط او التشابه في العبارات المستخدمة، كما احتسب الوسط الحسابي والانحراف المعياري للبيانات الواردة في الجدول(4)الذي يمثل الحقل الاول فيه مجموع المقالات التي تشترك عناوين الوثائق التي اشارت اليها كل مقالة باستخدام عبارات متشابهة تتمثل بعدد العناقيد في الحقل الثاني . ومن ثم تم تمثيل هذه البيانات في الشكلين (2، 3) حيث تظهر الحالات المتطرفة في طرفي المنحني لكل منها ففي الجدول (3) والشكل (2) يبدو ان هناك مقالتين لم يحصل ترابط فيها بين عبارات المستخلص وعبارات العناوين التي اشارت اليها في حين ان هناك ثلاث مقالات تجاوز فيها عدد العناقيد الستة وهو يبعد عن الوسط الحسابي ويقع على طرف المنحني .

مجموع المقالات	عدد العناقيد
2	0
11	1
16	2
20	3
13	4
4	5
3	6
69	21
69	21
الوسط الحسابي لعدد العناقيد = 3	
الانحراف المعياري = ؟	

الجدول رقم (3)

تمثيل مجموع المقالات وعدد العناقيد مع العناوين



الشكل (٢)

منحنى الترابط بين المستخلصات والوثائق المشار اليها

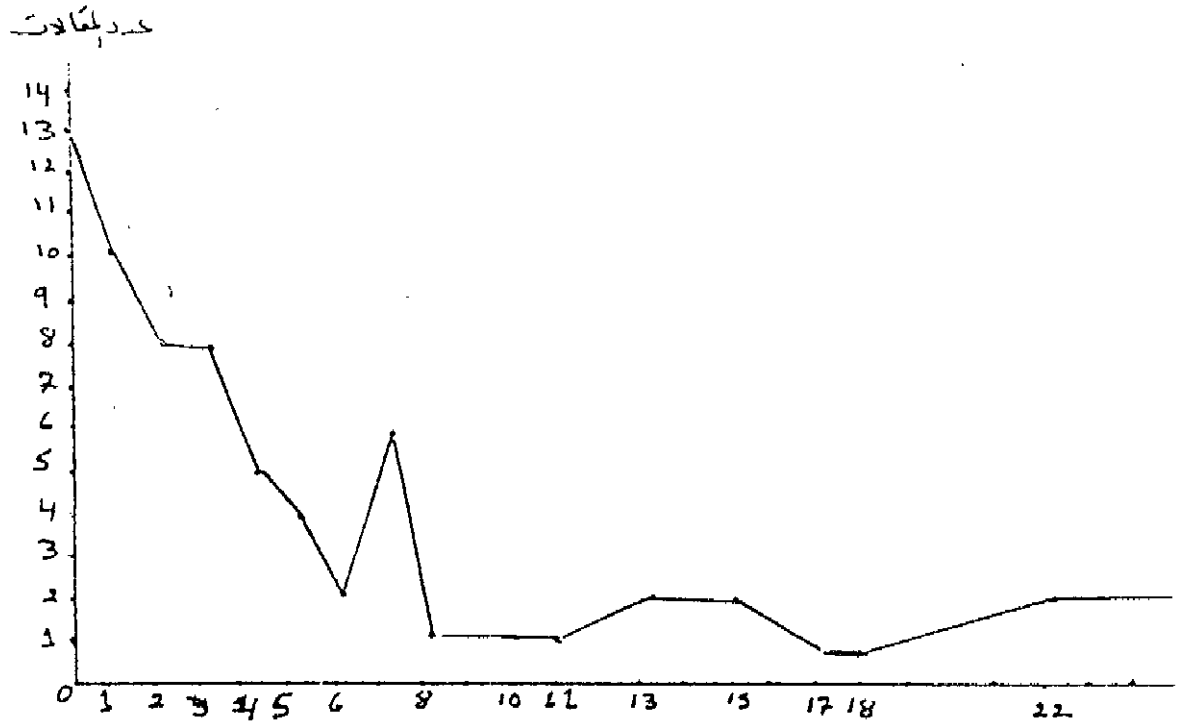
مجموع المقالات	عدد العناقيد	مجموع المقالات	عدد العناقيد
1	13	0	9	1	8
2	10	1	10	1	10
3	8	2	11	1	11
4	8	3	12	2	13
5	5	4	13	2	15
6	4	5	14	1	17
7	2	6	15	1	18
8	6	7	16	2	22
			17	2	30

مجموع المقالات =69 مجموع العناقيد = 172

الوسط الحسابي لعدد العناقيد =10

الانحراف المعياري =8

الجدول رقم (٧) تمثيل مجموع المقالات وعدد العناقيد مع العناوين



الشكل رقم (3)

منحنى الترابط بين عناوين الوثائق المشار اليها

التوصيات

١ ــ لغرض تقليص عدد العناقيد المتشتتة في المقال الواحد يقترح وضع عدد من الصيغ التي يعمل بموجبها على تعديل قائمة العبارات الدالة الحسرة وتتضمن هذه التعديلات الاتي :

ا ــ الانفاق على استخدام صيغة الجمع بدلا عن المفرد .

ب ــ الابتعاد عن المختصرات مثل (SDI) والتعبير عن العبارة كاملة ((Selective Dissemination of Information)

وان كانت مطولة .

ج ـ توحيد استخدام بعض العبارات التي تعطي مدلولا موضوعياً واحداً على سبيل المثال

- Technical Technological
- Aging Elderly
- Developing Countries / Third World / Less developed Countries

د ـ استخدام المضاف والمضاف اليه بدلا من الجملة .

Gathering of Information

تصبح

- Information gathering

٢ ـ لقد اقترح في دراسة سابقة (11) وتؤكد الاقتراح هنا الى استخدام النص الكامل للمقالة بدلا عن المستخلص في اختيار العبارات الدالة ومما يساعد على تحقيق هذا المقترح هو التطورات الواسعة في تقنيات الخزن والمعالجة الحديثة التي تسهل هذه العملية مثل استخدام الأقراص المكتنزة (D-ROM) وعليه فان شيوع استخدام هذه التقنية قد يدفع بطريقة اختيار العبارات الدالة من عنوان المقالة او مستخلصها نحو الزوال . اما بالنسبة الى عناوين الوثائق المشار اليها فانها استخدمت في الأساس لتغذية نظام الأسترجاع ولذا يبقى اعتمادها مفيدا لتحسين الأسترجاع ، وقد اكد على هذه الاهمية لعناوين الوثائق المشار اليها كليفلند (Cleveland) (12) حيث اوضح ان لعناوين الوثائق المشار اليها دوراً لايختلف عن النص الكامل للوثيقة في انجازها لتغذية النظام وتكامله بأرتباطها مع الوثيقة المصدرية .

1. Van Rijsbergen, C.J. Information Retrieval. 2nd ed. London: Butterworths, 1979.

2. Salton, G. and Megill, M.J. Introduction to Modern Information Retrieval .New York: mc graw Hill, 1983.

3. Ibid.

4. Can, Fazli and Ozkarahan, Esen A. "Similarity and Stability Analysis of the Tow Partitioning Type Clustering Algorithms," Jasis. 36: 1 (1985) 3–14.

5. Van Rijsbergen. OP. Cit.

6. Salton, G. Dynamic Information and Library Processing. Englwood. Cliffs, N.J.: Prentic-Hall, 1975.

7. Saracevic, T. "Relevance;a Review of and a Framework for the Thinking on the Notion in Information Science." JASIS. 26 (1975) 321–343.

8. Goffman, W. "An Indirect Method of Information Retrieval." Information Storage and Retrieval-4 (1969) 361–373'.

9. Kwok,K.L"A Probabilistic Theory of Indexing and Similarity Measure Based on Cited Citing Documents. JASIS. 36: 5 (1985) 342–351.

10. Philip, G; Smith, F.J. ;and Crookes, D. Voice Input/ Output Interface for Online Searching: Some Design and Human Factor Consideration. JIS. vol 14: 2 (1988) 93–93.

11. Kwok, K.L."A Probabilistic Theory of Indexing and Similarity Measure Based on Cited and Citing Documents. "JASIS. 35:5 (1985) 350.

12. Cleveland, D.B.; Cleveland, A.B.; and wise, O.B. "Less then Full text Indexing Using a Non–Boolean Searching Model." JASIS. 35 (1984) 19–28.