كفاءة وفاعلية استرجاع المعلومات باستخدام بعض خوارزميات مطابقة السلسة التقريبي

سهير عبد داؤد عيسى
مدرس مساعد/ قسم نظم معلومات
جامعة الموصل / كلية الإدارة والاقتصاد

أنهار خير الدين محمد
مدرس مساعد / قسم نظم معلومات
جامعة الموصل / كلية الإدارة والاقتصاد

الاستلام 15 / 05 / 2007
القبول 15 / 08 / 2007

ABSTRACT

The research aims at buliding integral computer database for sales ,by using six algorithms of approximate string matching with practicable example;soundex,information metaphone,longest common subsequence ,dice cofficient,levenshtein distance and fuzzy string matching .

The algorithm of matching is called (name similarity) to find the similarity between two strings depending on one kind of algorithm mentioned above as a limited digital value between (zero-one) representing the rate of information retrieval .while this rate is larger than 0.79 the matching between the two strings will be the best ,if the rate is smaller this will indicate a lack of conform of the strings. is addition to consider the precision of retrieve the intended records out of the existed records of the table.

الخلاصة

يهدف البحث إلى بناء قاعدة بيانات حاسوبية متكاملة للمبيعات، واستخدام ستة خوارزميات من خوارزميات مطابقة السلسلة التقريبي مع أمثلة تطبيقية وهي:- **(1)** Soundex ، **(2)** Double Metaphone **(3)** Longest Common Subsequence، **(4)** Dice Cofficient **(5)** Levenshtein Distance **(6)** فهرسة الصوت المضبب (Fuzzy String Matching). تم بناء خوارزمية مطابقة باسم (Name Similarity) من قبل الباحثين لإيجاد مقدار التشابه بين السلسلتين بالاعتماد على بعض خوارزميات المطابقة

أعلاه كقيمة رقمية محصورة بين (صفر-واحد) تمثل نسبة استرداد (اسـترجاع المعلومـات) فكلما كانت هذه النسبة أكبر من (0,79) كلما كان التطابق بين السلسلتين جيد، أما إذا كانـت هذه النسبة أقل فهذا يدل على عدم وجود تطابق بين السلسلتين، فضلاً عـن حـساب الدقـة (Precision) في استرجاع السجلات ذات العلاقة (المطلوبة) من مجموع السجلات الموجودة في الجدول.

المقدمة

تستخدم الأسماء بصورة متكررة معيار بحث لاسترداد المعلومات وللتعـرف وتمييـز الأنظمة المتكافئة أو المتماثلة، لذا فان الأسماء تقوم بدور هام في نظم المعلومات إذ أن بعض الأسماء في حالة التطبيق قد يحدث لها بعض الخلل منها خطأ في الهجاء، الألقـاب والكنيـة، الاختلافات الثقافية تعقد استرداد المعلومات المبنية على أساس الأسماء.

الدراسات السابقة:[5]

1- كانت هناك محاولة مبكرة من قبل رزل (Russel) وأودل (O'Dell) فـي تطـوير النظام العشري لفهرسة الصوت وذلك بتخصيص رمز صـوتي مـشترك لأسـماء وكلمات ذات اللفظ الصوتي المتشابه وهذا بدوره يجهز مقدرة بحث غير دقيقـة فـي استرداد المعلومات وذلك بموازنة طول النص المتغير مع الرموز العدديـة الحرفيـة (مكونة من عدد وحروف) ذات الطول الثابت حيث أن الدقـة والاسـتدعاء يكونـان لدرجة ما محدودتين لسبب أن عملية الترجمة تأخذ بنظر الاعتبار فقط حروفاً مفـردة مع استثناء الحروف الصامتة.
2- المؤلفين زامورا وإينجل قاما بالتغلب على بعـض التعقيـدات للنـصوص المعقـدة باستخدام الرسم الثلاثي لموازنة المساواة بين كلمتين.
3- المؤلف دامروا قام بعمل مقياس مشابه لحساب الفروقات بين كلمتين.
4- أما بالنسبة للمؤلفين سلكوا (Celko) وبفايفر (Pfeifer) فانهما يمنحان تحـسينات جوهرية للاقتراب الأصلي. اذ إن الطريقة المتبعة بواسطة سلكوا (Celko) تتغلـب على أخطاء الإدخال، الحذف، خروقات التعويض، تغير مواقع الحروف، في حين أن بفايفر (Pfeifer) قدم بوضوح دقة محسنة باستخدام أسلوب مبني على التماثل، إذ أن التشابه في إحراز النقاط خصص لتلائم مبني على عدد النقاط المشتركة بين أسـمين يحتويان على رموز صوتية مشتركة. إن معامل التـشابه عنـد بفـايفر (Pfeifer)

المرموز بـ (δ) لحالة الاسم الموسوم بالنقاط ذات الوزن (α) ونتائج اسم آخـر ذات الوزن (β) يكون :-

$$\delta = \frac{\alpha \cap \beta}{\alpha \cup \beta}$$

أهداف البحث

1. بناء قاعدة بيانات حاسوبية متكاملة خاصة بالمبيعات.
2. استخدام ستة خوارزميات من خوارزميات مطابقة السلسلة التقريبي.
3. بناء (تصميم) خوارزمية باسم **(Name Similarity)** والتي تـستخدم بعـض مـن خوارزميات مطابقة السلسلة التقريبي في بناءها لإيجـاد مقـدار التـشابه (Similarity Score) بين السلسلتين، ومن ثم إيجاد نسبة الدقة لتحسين الاستدعاء المبني على أسـاس الأسماء في استرداد المعلومات بدون الخلل في الدقة.

1. المبيعات

المبيعات: كلمة مشتقة من كلمة بيع، والمبيعات هي نتيجة لعملية البيع وعندما تتم عمليـة البيع نستطيع أن نقول أن المبيعات قد تحققت.

الكثير من الناس يخلط بين التسويق والمبيعات، **فالمبيعات** هي جزء من التـسويق وهـي تبادل المعلومات لإقناع المشتري لشراء شيء قد يكون منتجاً ملموساً أو خدمة أو فكرة والتي في النهاية ترضي احتياج المشتري.

هناك شرطان لتحقيق عملية البيع (المادي والمعنوي)، **فالمادي** منها هو ما نقبـضه مقابـل السلعة التي نبيعها، أما **المعنوي** هو ما تم من موافقة المشتري عليه بالرضى والقبول بعمليـة البيع. **وهذان الشرطان الأساسيان لعملية البيع لهما أسس يجب العمل على تحقيقهـا لكـي يكتمل الشرط :-**

- **فالمادي :-** من أسس تحقيق العائد من المبيعات هي إيجاد المنافسة للـسلعة المـراد بيعها فيكون سعرها منافساً لأسعار السوق ولمثيلاتها من السلع في الأسواق، و كذلك توفر الجودة في المنتج بما يماثل أيضا المنتجات المتوفرة في الأسواق. إن المـردود المـادي يتطلب من البائع صرف تكاليف قد تكون في بعض الأحيان مكلفة للغاية وذلـك لهـدف بسيط نوعا ما ولكنه مطلوب بشكل رئيس وهو الدعاية للمنتج التي يكون هـدفها الوحيـد (تعريف المشترين بوجود المنتج ذو الجودة كذا وكذا وبمواصفات أيـضا معينـة لأكبـر شريحة من المستهلكين).

- **أما بالنسبة للمعنوي :-** لكي توصل القناعة الفعلية للمشتري بشراء منتجك يجـب أن تحقق له الرغبة في الشراء وتزكي منتجك له بفكرة بسيطة قد تخرج أحيانا إلى نطاق شد الانتباه من خلال دعاية تلفزيونية أو حتى دعاية بإحدى الجرائد تلفت انتباهه للمنتج والتي من خلالها يبني فكرة مميزة تزكي منتجك عن باقي المنتجات التي كان يـستخدمها مـن ناحية النوعية أو الجودة أو حتى من خلال تعميم المنتج نفسه أو الخدمة المراد بيعها. **[2]**

2. قاعدة البيانات

هو عبارة عن مجموعة من الجداول (Tables) يتألف كل جدول مـن أعمـدة أو حقـول (Column Fields) وسجلات أو صفوف (Records) ويؤدي تقاطع العمود مع الصف إلى تقديم الخلية (Cell) حيث يتم تخزين معلومة واحدة فقط في الخلية الواحدة .

وهذا تعريف قاعدة البيانات من الناحية الفيزيائية أما تعريف قاعدة البيانـات مـن الناحيـة المنطقية إذا صح التعبير فهو المستودع الذي يضم ويحوي جميع المعلومات عن منظمـة أو شركة أو تنظيم ... الخ مهما كان حجم هذا التنظيم أو نوع المعلومات .

وتأتي أهمية قاعدة البيانات الإلكترونية لما تحققه من سهولة التخزين وسرعة في التوصل إلى أي بحث عن سجل ما وإجراء المقارنات والإحصائيات ... الخ**.[1]**

3. نظم استرجاع المعلومات

نظم استرجاع المعلومات هي أحد النظم الفرعية في نظام المعلومات، وظيفته الأساسـية هو اختزان واسترجاع المعلومات وفقا لاحتياجات مجتمع المستفيدين سواء تم داخل المكتبات أو خارجها فانه في الحقيقة وفي اغلب الحالات مجرد وسيلة أو أداة يصل من خلالها الباحث والمستفيد إلى وعاء المعلومات التي يبحث عنه ، فالذي يختزن عن هذه الأوعيـة بواسـطة الحاسب الإلكتروني في (بنك أو نظام معلومات) ليس إلا بيانات محدودة عـن كـل وعـاء، كعنوانه، وتاريخه والمسؤول عن محتواه الفكري ونشره، وعدد صفحاته أو أوراقه، ورؤوس الموضوعات فيه [11].

ويمكن تقسيم تلك النظم إلى :-

- **نظام استرجاع الحقائق Fact Retrieve System**

وهي تلك النظم التي يرجع إليها الباحث للحصول على المعلومات بشكل مباشر وتتمثل هذه النظم في مختلف أنواع المراجع:-

ككتب الحقائق، والكتب السنوية والإحصائية، والأدلة، وكتب التـراجم، ومعـاجم البلـدان، ومعاجم الموسوعات في شكلها المقروء آلياً (المحسب) التي يمكن أن نطلـق عيهـا بنـوك المعلومات (Data Bank) أو مصارف المعلومات.

- **نظم اسـترجاع الوثـائق Document Retrieve System أو نظـم اسـترجاع (الببلوجرافية) أو نظم استرجاع بدائل الوثائق**

وهي تلك النظم التي يرجع إليها الباحث للحصول على قوائم بالوثائق المتعلقة بموضوع معين، أو مؤلف ما، أو فترة زمنية محددة ... الخ.

ويتم الاستعانة هنا (ببلوجرافية) سواء من شكلها التقليدي أو المقروء آليا (أو ما يعـرف الآن بمراصد البيانات (database) للإجابة على استفسارات المستفيد).

4. خوارزميات مطابقة السلسلة التقريبي

4-1 خوارزمية Soundex

تعتمد هذه الخوارزمية على لفظ اللغة الإنكليزية وهي الأقدم **(نشرت عام 1918)**، حيث تعمل بالشكل الآتي [4]:

1) تحويل الحروف الصغيرة إلى كبيرة، مع إلغاء علامة الاقتباس إن وجدت.

2) الحفاظ على أول حرف من السلسلة وتحول باقي أحرف السلسلة إلى أرقام بالاعتماد على الجدول الآتي:-

جدول (1): تحويل الأحرف إلى أرقام الخاص بخوارزمية الـ (Soundex)

الأحـــــرف	الرقـــم
A ,E , I , O , U , H , W , **Y** (على أن لا تكون أول حرف من الكلمة)	**0**
B , F , P , V	**1**
C , G , J , K, Q , S , X , Z	**2**
D , T	**3**
L	**4**
M , N	**5**
R	**6**

3) الأرقام المتكررة سوف تحذف وتستبدل برقم واحد مثل $(3 \rightarrow 333)$.

4) حذف جميع الأصفار (التي عوضت بدل حروف العلـة أو (H,Y,W) والموجـودة فـي السطر الأول من الجدول) من السلسة الناتجة بعد الخطوة 3 .

5) السلسلة الناتجة تتكون من الحرف الأول من السلسة الأصلية مضاف لها ثلاثة أرقام، حيث أن السلسة التي هي اكبر من أربعة رموز سوف تقطع إلى أربعة، بينما القصيرة التي هي أقل من أربعة رموز سوف يكمل بأصفار مثال :-

Peter يحول إلى 'P360' أما الكلمة Chirsten يحول إلى'C623'

2-4 خوارزمية Double Metaphone

نشرت هذه الخوارزمية في ديسمبر عام 1990 في مجلة للحاسبات الإلكترونية وقـد طورت هذه الخوارزمية من قبل لورنس فيلبس (Lawrence Philips) ونشرت في مجلـة **C/C++** في حزيران من عام 2000 كجزء من الخوارزميات المعروفة بـ "مطابقة الصوت" أو "التشفير الصوتي" وهي كتبت بدلاً من خوارزمية Soundex هـذه الخوارزميـة تحـاول اكتشاف أصوات بديلة لنفس الكلمة، أو علاقات بين الكلمات على سبيل المثال عند تطبيق هذه الخوارزمية يلاحظ وجود علاقة صوتية قوية بين كلمة Nilson و Nelson حيث لها نفـس الرمز الصوتي ويلاحظ لا يوجد علاقة صوتية بين Nelson و Adem . هذه الخوارزميـة صممت في البداية لتشفير أسماء اللغة الإنكليزية الأمريكية باستخدام أشـهر قواعـد الـتلفظ الإنكليزية التي لا تستطيع خوارزمية الـ Soundex أن تغطيها.**[3]**

هذه الخوارزمية تجعل الكلمة بين **(1 – 4)** حروف باستخدام ابسط القواعد الـصوتية للتلفظ الإنكليزي المعتاد إذ تأخذ في الحسبان أن هذه الكلمات يمكن أن تأخذ أكثر مـن تلفـظ مقبول واحد حيث تحسب هذه الخوارزمية التشفير الأولي والتشفير الثانوي للكلمـة المعطـاة للإشارة إلى أن لكليتهما على الأغلب نفس التلفظ، من هنا جاءت التسمية Double .

هذه الخوارزمية تتجاهل أحرف العلة التي تأتي بعد أول حرف، فقط أحرف العلة التي تأتي في بداية الكلمة تبقى حيث تقلل هذه الخوارزمية الحروف الصوتية إلـى (16) صـوت ساكن هي **(B,X,S,K,J,T,F,H,L,M,N,P,R,O,W,Y)** حيث "**sh**" تمثل بـ X وصفر بدل "**th**" . الجدول (2) ، (3) يوضحان القواعد المطبقة مع بعض الشواذ (الاسـتثناءات). [8] وكما يلي :-

جدول (2): قواعد تشفير Metaphone

| الحرف | الرمز | تعليـــــــ

صامتة إذا لا تتبع بحرف علة		W
صامتة إذا تتبع بحرف علة	W	
	KS	X
صامتة إذا لا تتبع بحرف علة		Y
صامتة إذا تتبع بحرف علة	Y	
	S	Z

جدول (3): استثناءات تشفير Metaphone

احذف أول حرف	إذا البداية "wr-" , "ae-" , "pm-" , "gn-" , "kn-"
تغير إلى s	إذا البداية "x"
تغير إلى w	إذا البداية "wh-"

3-4 خوارزمية Longest Common Subsequence (LCS)

تستخدم هذه الخوارزمية لإيجاد أطول سلسلة متشابهة بين سلسلتين بمقارنة السلسلتين ومعرفة عدد الحروف المضافة في سلسلة والمحذوفة من سلسلة أخرى حيث يتم إيجاد السلسلة الفرعية التي تمثل أطول سلسلة متشابهة بين السلسلتين [9].

فمثلاً :- إذا كان لدينا:

السلسلة الأولى هي :- **X (a b c d a c e)**

والسلسلة الثانية هي :- **Y (b a d c a b e)**

حيث أن السلسلة الأولى يمكن تمثيلها بـ

X= x1 , x2 , x3 , ……. ,xm

والسلسلة الثانية يمكن تمثيلها بـ

Y= y1 , y2 , y3 ,…….,yn

Xi

X			**i**		

Yj

Y			**J**		

حيث $0 <= i <= m$ و $0 <= j <= n$

لإيجاد أطول سلسلة فرعية للسلسلتين LCSS(X1i, Y2j) تتم عـن طريـق الخوارزميـة التالية[9]:-

```
LCS- Length (X,Y)

m ←——— Length[X]
n ←——— Length[Y]
for i ←———— 1 To m
do c[i,0] = 0
for j ←———— 1 To n
do c[0,j] = 0
for i ←————— 1 To m
do for j ←————— 1 To n
if xi = yj  Then
        c[i,j] = c[i-1,j-1] + 1
        b[i,j] = "↖"
LCS – length (x,y) continue
if Xi <> yj then
if c[i-1,j] >= c[i, j-1] Then
  c[i,j] = c[i-1,j]
  b[i,j] = "↑"
else c[i,j] = c[i,j-1]
   b[i,j] = "←"
return c and b
```

ومن هنا يتضح أن تمثيل السلسلتين يكون على شكل مصفوفة موضحة في الجدول (4) فـي خلاله نصل إلى أفضل سلسلة فرعية والتي يتم تمثيلها إبتداء من الزاويـة **الـسفلى اليمنـى** وباتجاه **الأعلى** من الجدول (4).

جدول (4) : نتائج تطبيق خوارزمية (LCS) .

	a.	b.	c.	d.	a.	c.	e.
b.	0	1	1	1	1	1	1
a.	1	**1**	1	1	2	2	2
d.	0	1	1	**2**	2	2	2
c.	1	1	2	2	2	3	3
a.	1	1	2	2	**3**	3	3
b.	0	2	2	2	3	3	3
e.	0	2	2	2	3	3	**4**

حيث أن أفضل سلسلة فرعية يمكن معرفتها من الجدول هي **(b d a e)** وأن عدد الحروف التي تمثل أطول سلسلة متشابهة من السلسلتين هي (4).

Length LCSS = 4

وأن المسافة المتغيرة تحسب بعدد العناصر المحذوفة وعدد العناصر المضافة إلى السلسلتين .

- edit distance = m + n – 2L
  = 7 + 7 – 2 * 4 = 6
- edit distance = 3(remove) + 3(add) = 6

حيث أن :-

m :طول السلسلة الأولى .

n :طول السلسلة الثانية .

أما الوقت المستغرق لملئ خلية من المصفوفة هو O(1)

وبما أن عدد مداخل المصفوفة هو m * n

- **زمن التنفيذ يمثل** O (m * n)

4-4 خوارزمية Dice Cofficient

تستخدم هذه الخوارزمية طريقة الأوزان (n-gram) وأحياناً يطلق عليها (Q– gram)

يمثل طول الوزن **Q أو n**

n–gram : monograms , bigrams , trigrams ; etc .

مثال :-

bigram for "Nelson"
%N Ne el ls so on n #

حيث :-

- **D** هي **Dice Cofficient**
- **SB(Shared Bigrams)** هي **المشترك من الأوزان**
- **TBg1(Total Bigrams 1)** **المجموع الكلي للوزن بالنسبة للسلسة الأولى.**
- **TBg2(Total Bigrams 2)** **المجموع الكلي للوزن بالنسبة للسلسة الثانية,**

فيصبح القانون:-

D = (2 * SB) / (TBg1 + TBg2)

فإذا أخذنا الكلمتين:-

- **Nelson**
- **Neilsen**

فعند تطبيق الوزن (n – gram) على السلسلة الأولى والسلسة الثانية نجد أن :-

SB هو **(n , N , Ls , Ne)**

TBg1 هو **(7)**

TBg2 هو **(8)**

D = 2 * 4 / (15) = 0.53

وبين كلمة **Nelson** و **Nelsen** **فإن :-**

D = 2 * 5 / (14) = 0.714

- أحسن **Dice Cofficient** ممكن أن تكون قيمتها أكبر من **(0.2)** وإذا كانت أقل من هذه النسبة ترفض [3]

5-4 خوارزمية Levenshtein Distance

سميت بهذا الاسم من قبل العالم الروسي Vladimir Levensth الذي اخترعها سنة 1965، حيث تستخدم الخوارزمية مقياساً للتشابه بين سلسلتين، السلسلة الأولى تسمى المصدر **(S)** والسلسلة الثانية تسمى الهدف **(t)** وذلك بحساب المسافة التي تمثل عدد **(المحذوفات والإدخال أو البدائل)** التي تمثل التغييرات، المسافة يرمز لها بـ **(d)**، ويمكن تمثيل الخوارزمية بالخطوات الآتية [7]:-

1- نجعل n تمثل طول السلسلة الأصلية S .

نجعل m تمثل طول السلسلة الهدف t .

إذا n = 0 نرجع m ويتم الخروج من الخوارزمية .

إذا m = 0 نرجع n ويتم الخروج من الخوارزمية .

تكوين مصفوفة من m … 0 من الصفوف من n … 0 من الأعمدة .

2- تهيئة أول صف من n … 0 وتهيئة أول عمود من m … 0 .

3- فحص كل حرف من السلسة (Si=1 … n) .

4- فحص كل حرف من السلسة (tj=1 … m) .

5- إذا S[i] تساوي t[j] تكون الكلفة (cost) مساوية للصفر .

إذا S[i] لا تساوي t[j] تكون الكلفة (cost) مساوية للواحد .

6- نجعل الخلية d [i, j] للمصفوفة مساوية لأصغر (لأقل) :-

I. الخلية الأعلى زائد واحد 1 (d [i - 1, j]+1) .

II. الخلية المباشرة إلى اليسار زائد واحد (d [i ,j -1]+1) .

III. الخلية القطرية إلى الأعلى والى اليسار زائد الكلفة (d [i - 1 , j -1]+cost) .

7- إذا كان i<=n و j<=m ، اذهب إلى الخطوة 3 .

8- تحسب المسافة ، وتخزن في الخلية d.

4-6 فهرسة الصوت المضبب

4-6-1 قواعد التحويل

هناك عدة قواعد تحكم عملية التحويل ويتم وصفها :-

I. الإزالة

- **إزالة الحرف الساكن**
- الإزالة عند وجود حرفين ساكنين

عند تكرار حرفين ساكنين متعاقبين، يزال حرف واحد .

- **إزالة حرف العلة**

حروف العلة التي تأتي في بداية الكلمة يحتفظ بها أما حروف العلة التي لا تأتي في بداية الكلمة تزال .

- **إزالة الحروف في نهاية الكلمة**

تهمل بعض الحروف معتمدا على **"n-gram"** ((وزن تحديد الهوية))، الذي يأتي في نهاية الكلمة.

II. الاحتفاظ

- **الحروف التي تأتي في بداية الكلمة**

إن الرموز والرموز الحرفية التي تأتي في بداية الكلمة عادة يحتفظ بهـا بـشكل حرف هجاء عادية، وإنتاج رموز رقمية للمواقع التي ليست في بداية الكلمة.

4-6-2 خوارزمية فهرسة الصوت المضبب[5]

1-احتفظ بالحرف الأول من السلسلة .

2-استخدام قواعد التعويض العائدة لـ **"n-gram"** وكما موضح في الجدول (5)

جدول (5): تعويض وزن تحديد الهوية (n-gram) بحرف أو حروف

وزن تحديد الصوت n-gram	حروف في البداية	الحروف اللاحقة	التعويض
CA			KA
CC , CK			KK
CE			SE
CH		KK	
CHL , CL			KL
CHR , CR			KR
CI			SI
CO			KO
CS , CZ , TS , TZ	SS		
CU			KU
CY			SY
DG			GG
GH			HH
GN	NN		
HR , WR	RR		
HW	WW		
KN , NG	NN		
MAC , MC			MK
NST			NSS
NT		TT	
PF , PH			FF
RT , RDT		RR	
SCH			SSS
TIO , TIA			SIO
TCH			CHH

3- طبق قواعد التحويل على السلسلة.

4- تحويل الحروف إلى رموز رقمية (التحويل الرمزي)، وكما موضح في الجدول (6).

جدول (6) : تعويضات الحروف بأرقام

الحروف	التحويل الرمزي بطريقة Fuzzy
A , E , I , O , U	
B , P	1
C	9
D , T	3
F , V	1
G , J , K , Q	**7**
H , W , Y	
L	4

الحروف	التحويل الرمزي بطريقة Fuzzy
M , N	5
R	6
S , Z	9
X	7

5- احذف جميع الحروف للسلسلة مع الإبقاء على الحرف الأول.

6- رجع أول أربعة حروف أو رموز، من جهة اليمين احشر أصفار إلى السلسلة إذا كان طول السلسلة أقل من (4 Byte).

5-خوارزمية المطابقة(التشابه) (Name Similarity)

تم تصميم خوارزمية مطابقة * باسم (Name Similarity) تقيس درجة التشابه بين السلسلتين بالاعتماد على خوارزميات المطابقة (Levenshtein Distance، Longest Common Subsequence (LCS)، Dice Cofficient) التي تعتمد على (Bigram)وكما موضحة في الخوارزمية أدناه :-

```
1  ms ←———— Length(String1) : mk ←———— Length(String2)
2  Lc ←———— LCS(String1, String2)
3  Le ←———— Levenshtein (String1, String2)
4  D ←———— Dice (String1, String2)
5  B ←———— Bigram(String1) ,  B1 ←———— Bigram(String2)
6  BB ←———— B + B1
7  MaxLen ←———— ms ,  MinLen ←———— ms
8  if mk >  MaxLen Then MaxLen = mk
9  if mk <  MinLen  Then MinLen = mk
10 if MaxLen = MinLen and Le=0 Then Score = 1
   Exit Algorithm
11 if (ms > mk or mk < ms) and Le >=5 and Lc >10 Then
      Score = ((MinLen + Lc + (Lc-Le)) / (MinLen + MaxLen))
      +(D/BB)
   Exit Algorithm
12 if (ms < mk or ms > mk) and Le <=5 Then
      Score = (MinLen / MaxLen) +(D/BB)
   Exit Algorithm
13 if ms = mk and Le >0 and Lc >0 Then
      Score = ((MinLen + Lc) / (MinLen + MaxLen)) + (D/BB)
   Exit Algorithm
14 Return Score
```

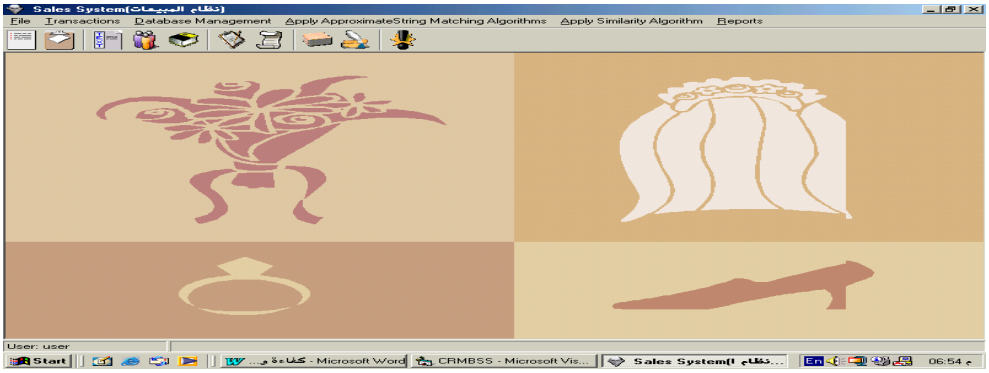
* هذه الخوارزمية من إعداد الباحثين

6-ملفات النظام :-

تم تصميم قاعدة البيانات (Sales.Mdb) باستخدام برمجيات (Access) وربطها مع فيجوال بيسك بإصداره السادس) والتي تحتوي على العديد من الجداول موضحة في الملحق (1).

7-تصفح النظام

عند النقر على أيقونة النظام تظهر الشاشة الرئيسة الموضحة في الشكل (1) والتي تحتوي على الأوامر التالية :-



شكل (1) الشاشة الرئيسة للنظام

1- أمر File :-

يتكون هذا الأمر من أمر ثانوي واحد وهو :-

- **أمر Exit :-** يتم إغلاق جميع الملفات المفتوحة والخروج من النظام.

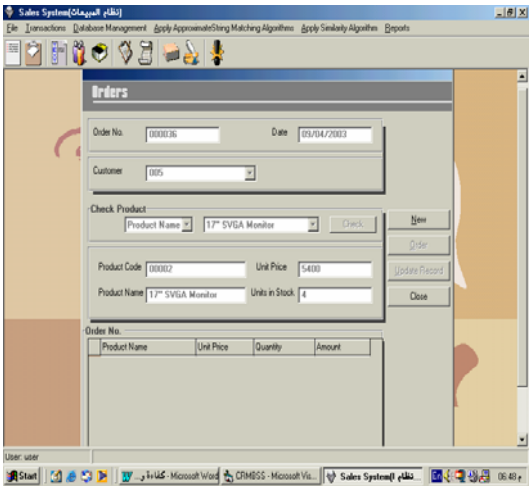
2- أمر Transactions :-

يتكون هذا الأمر من أمرين ثانويين :-

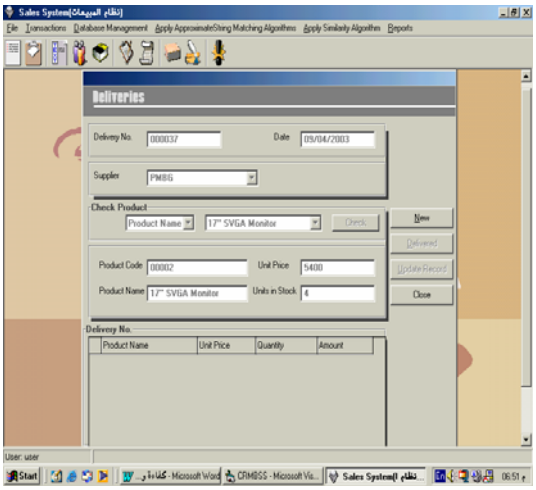
1- **أمر Orders :-** يقوم بإتمام إجراءات الطلبات، وكما موضح في الشكل (2)

2- **أمر Deliveries :-** يقوم بإجراءات عملية التسليم للمبيعات المطلوبة في الطلبات، وكما

موضح في الشكل (3).



شكل (2) شاشة الطلبات

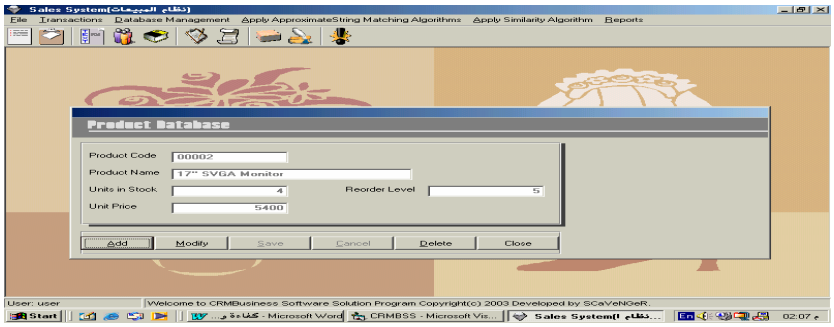


شكل (3) شاشة تسليم الطلبات

3- أمر Data Management :-

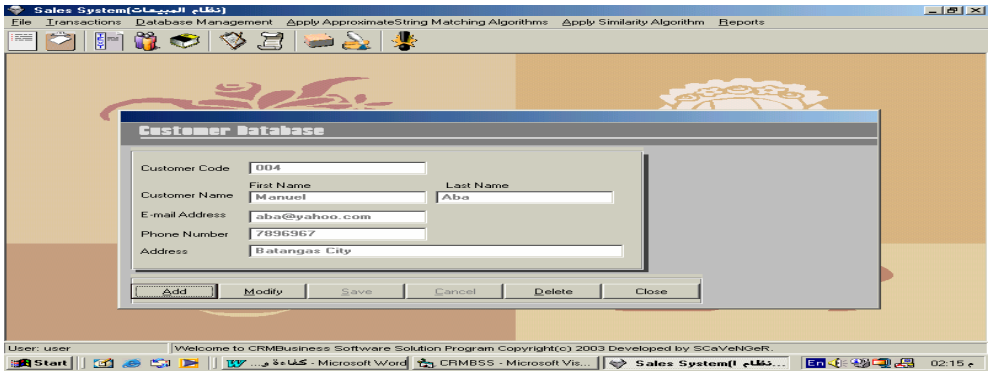
يتكون من ثلاثة أوامر ثانوية تتعامل مع الملفات الرئيسة لنظام المبيعات وكما يأتي :-

1- أمر Product Database :- يقوم بإدارة ملف المنتج (Product) فضلاً عن إجراء عمليات الحذف، الإضافة،... الخ على بيانات هذا الملف وكما موضح في الشكل (4) .



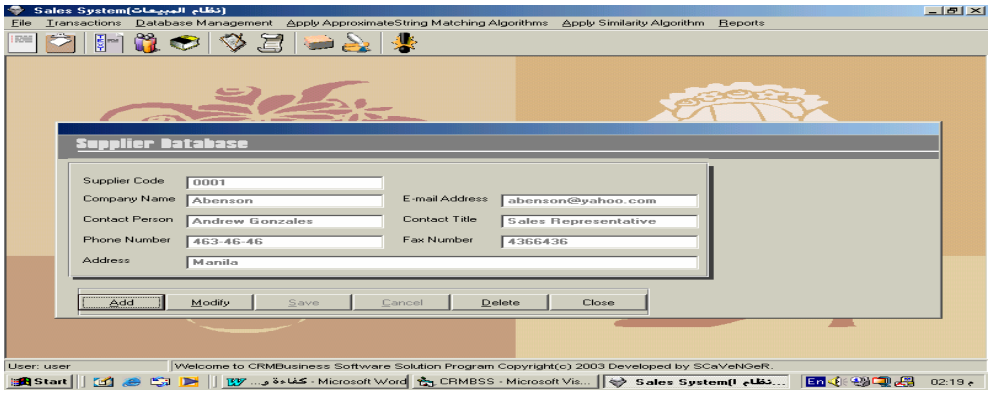
شكل (4) شاشة المنتج

2-أمر Customer Database :- يقوم بإدارة ملف الزبائن (customers) فضلاً عن إجراء عمليات الحذف، الإضافة، ... الخ على بيانات هذا الملف وكما موضح في الشكل (5).



شكل(5) شاشة الزبون

3- أمـر Supplier Database :- يقوم بإدارة ملف المجهزين (Suppliers) فـضلاً عن إجراء عمليات الحذف، الإضافة، ... الخ على بيانات هذا الملف وكما موضح في الشكل

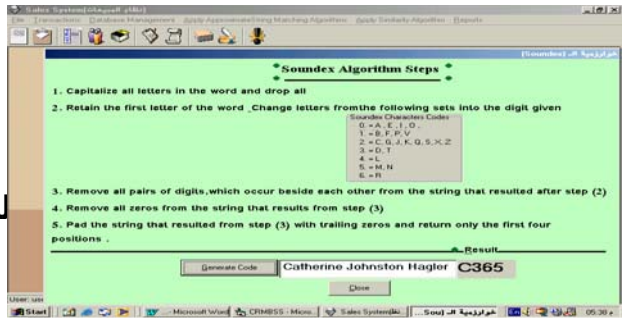


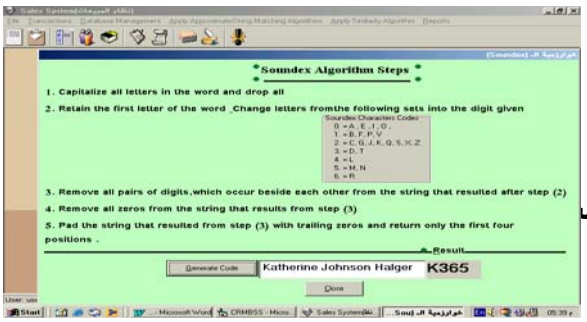
.(6)

4- أمر Apply Approximate string matching Algorithms

يتكون هذا الأمر من خمسة أوامر ثانوية وكما يأتي :-

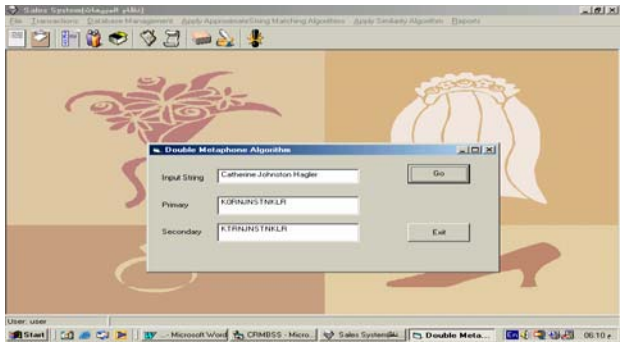
1) أمر Soundex Algorithm:- عند الكبس على هذا الأمر سوف تظهـر الـشاشة الخاصة باستخدام هذه الخوارزمية، وكما موضح في الشكل (7) و (8) .



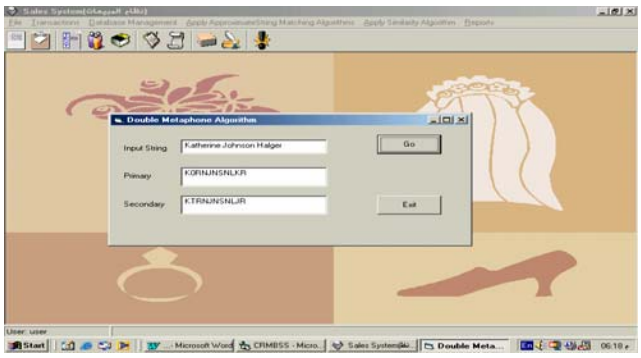


شكل(7)شاشة خوارزمية الـ(Soundex) للسلسلة الأولى **شكل(8)شاشة خوارزمية الـ(Soundex) للسلسلة الثانية**

2) أمر Double Metaphone Algorithm:- عند الكبس على هذا الأمر سوف تظهر الشاشة الخاصة باستخدام هذه الخوارزمية، وكما موضح في الشكل (9) ، (10) .

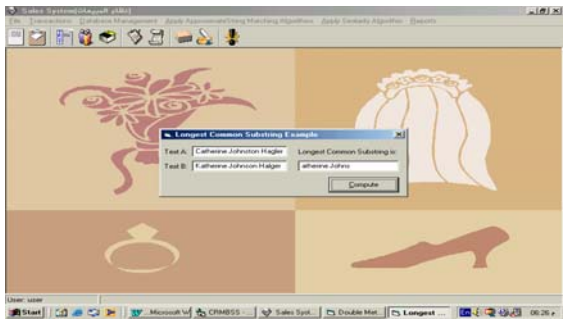


شكل(9) شاشة خوارزمية الـ(Double Metaphone) للسلسلة الأولى

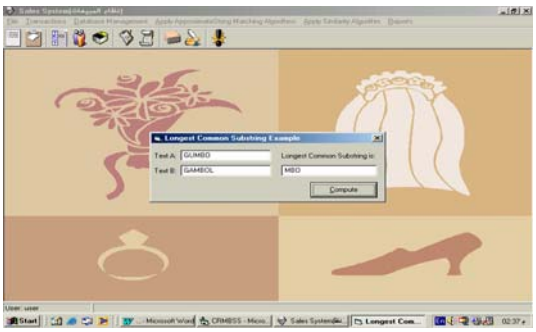


شكل(10) شاشة خوارزمية الـ(Double Metaphone) للسلسلة الثانية

3) أمر Longest Common Subsequence (LCS)) Algorithm):- عند الكبس على هذا الأمر سوف تظهر الشاشة الخاصة باستخدام هذه الخوارزمية، وكما موضح في الشكل (11) و (12) .

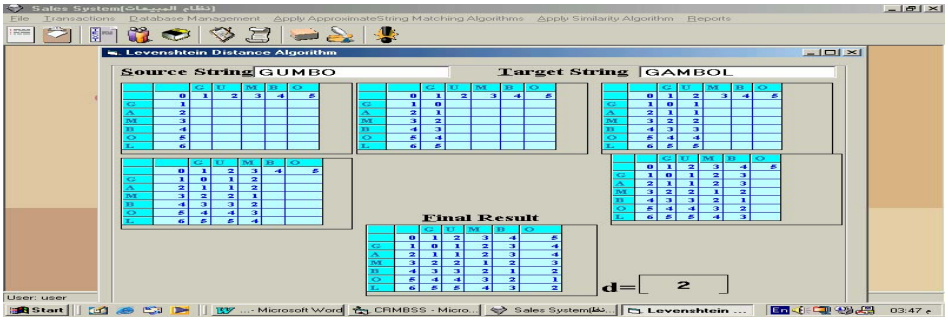


شكل(11) شاشة خوارزمية الـ (Longest Common Subsequence (LCS)) للمثال الأول



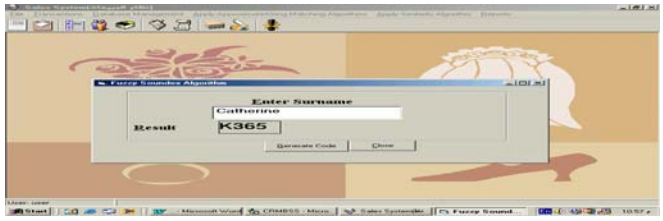
شكل(12) شاشة خوارزمية الـ (Longest Common Subsequence (LCS)) للمثال الثاني

4) أمر Levenshtein Distance:- عند الكبس على هذا الأمر تظهر الشاشة الخاصة بهذه الخوارزمية وكما موضح في الشكل (13).

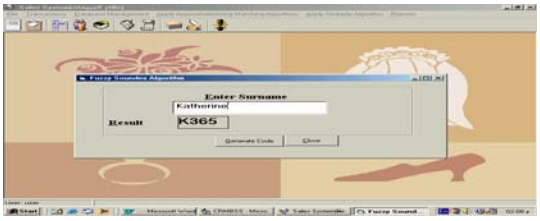


شكل (13) الشاشة الخاصة بخوارزمية Levenshtein Distance

5) أمر Fuzzy Soundex Algorithm:- عند الكبس على هذا الأمـر سـوف تظهـر الشاشة الخاصة بهذه الخوارزمية، وكما موضحة في الشكل (14) و(15) .



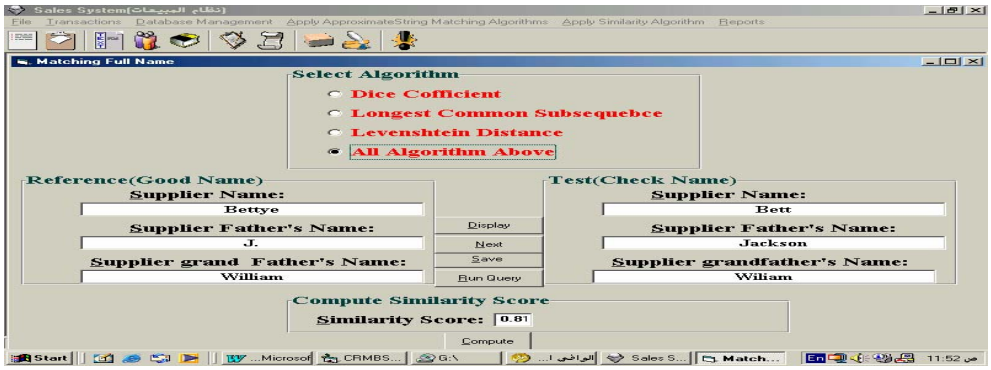
شكل(14) شاشة خوارزمية الـ (Fuzzy Soundex Algorithm) للسلسلة الأولى



شكل (15) شاشة خوارزمية الـ (Fuzzy Soundex Algorithm) للسلسة الثانية

5- أمر *Apply Similarity Algorithm*

تستخدم خوارزمية المطابقة (Name Similarity) والموضحة في الفقرة (5) لحـساب درجة التشابه بين سلسلتين حيث تحسب درجة المطابقة بين السلسلتين بشكل تقريبي، حيث أن نسبة التطابق المحسوبة تكون محصورة بين (صفر-1)، فإذا كانت نسبة التطابق أكبر من أو مساوية لـ (0,79) تكون مقبولة، أما إذا كانت هذه النسبة أقل من (0,79) تعتبر نسبة غيـر مقبولة **(أي لا يوجد تطابق بين السلسلتين)** .



شكل (16) شاشة عملية المطابقة

6- أمر *Reports* :-

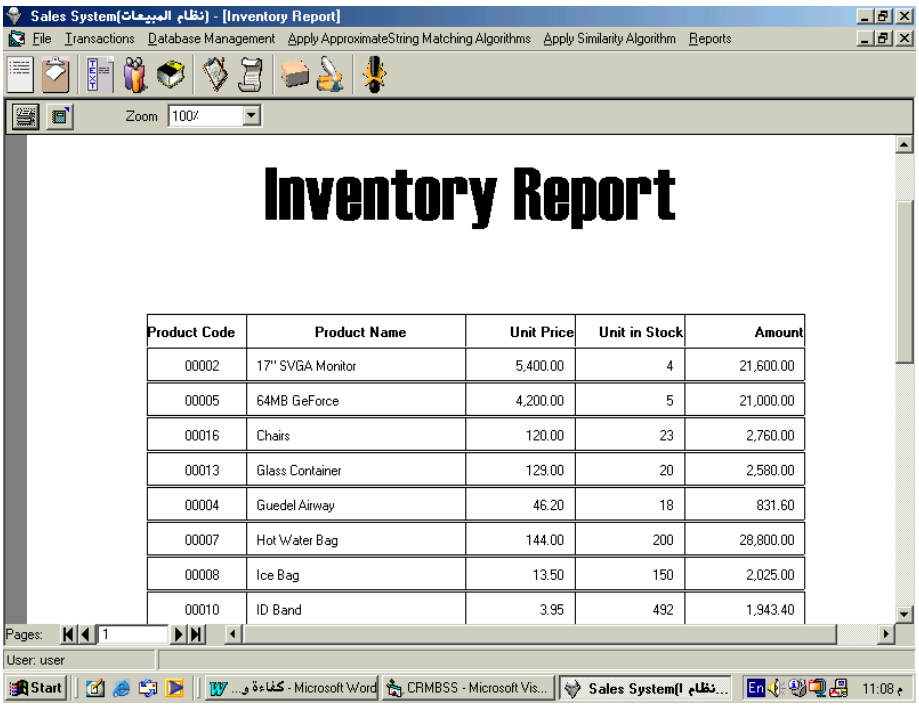
يتكون هذا الأمر من أمرين ثانويين :-

1- **أمر Sales Reports :-** يقوم بإصدار تقرير خاص بالمبيعات، وكما موضـح فـي الشكل (17)

Order No.	Product Name	Unit Price	Quantity	Amount
000036	Hot Magic Bag	120.00	2	240.00
000036	Hot Magic Bag	120.00	4	480.00
000036	Chairs	120.00	12	1,440.00
000038	Glucomen Test Strip	46.20	2	92.40
000038	Glucomen Test Strip	46.20	6	277.20
000039	ID Bracelet Pink & Blue	2.60	2	5.20
000041	Seagate 10GB Hardisk	4,500.00	5	22,500.00
000042	Hot Magic Bag	120.00	11	1,320.00
000042	Hot Water Bag	144.00	10	1,440.00

شكل(17) تقرير خاص بالمبيعات

2- **أمر Inventory Reports :-** يقوم بإصدار تقرير خاص بجرد المنتوجـات، وكمـا موضح في الشكل (18).



Product Code	Product Name	Unit Price	Unit in Stock	Amount
00002	17" SVGA Monitor	5,400.00	4	21,600.00
00005	64MB GeForce	4,200.00	5	21,000.00
00016	Chairs	120.00	23	2,760.00
00013	Glass Container	129.00	20	2,580.00
00004	Guedel Airway	46.20	18	831.60
00007	Hot Water Bag	144.00	200	28,800.00
00008	Ice Bag	13.50	150	2,025.00
00010	ID Band	3.95	492	1,943.40

شكل (18) تقرير خاص بجرد المنتوجات

الاستنتاجات :

1- استخدام مطابقة السلسة التقريبي مع مجموعة البيانات الكبيرة مرغوب ومطلوب للمشاكل التي يمكن أن تنشأ والمرتبطة بمجموعة النص والبيانات.

2- من خلال قاعدة البيانات التي تم تكوينها باسم **Sales** والتي تضم (17295) سجل يضم (اسم المجهز، اسم الأب، اسم الجد) تـم تكـوين نـسخة أخـرى مـن الجـدول باسـم (SupplierTest) يحتوي على معلومات مماثلة مع اختلافات في الأسماء نتيجة الطباعة أو التهجئة أو الألقاب ويضم (1200) سجل، ولغرض قياس درجة التشابه بين الأسـماء المتطابقة بين الجدولين ونسبة التطابق **(التشابه)** حيث إذا كانت نسبة التشابه أكبـر مـن (0,79) فهذا يعني أن نسبة التشابه بين السلسلتين جيـدة وتـضاف إلـى جـدول الـ (Result)، أما إذا كانت النسبة أقل من (0,79) فهذا يعني أنـه لا يوجـد تطـابق بـين السلسلتين وتوضع في جدول آخر لعدم التطابق باسم (Unmatch File). وبعـد تنفيـذ الاستعلام (استرجاع المعلومات المبني على الأسماء) على السجلات الموجودة في جداول التطابق ونتائجه (اعتماداً على نتائج تطبيق خوارزمية (Name Similarity)).

كانت نتائج استعلام المطابقة التقريبي كما يأتي :-

- 1322 سجلاً مخزونة في سجلات جدول الـ (Result) مع نسبة التطابق.
- مجموع المطابقة التقريبي هو 1188.
- نسبة الدقة هي 99% .

تم حساب نسبة الدقة بالاعتماد على القانون التالي :-

نسبة الدقة = (عدد السجلات المسترجعة ذات العلاقة / العدد الكلـي للـسجلات)* 100 [6] .

نسبة الدقة = (1188 / 1200) * 100 =0,99

- عدد الأسماء غير المتطابقة 12 .
- نسبة عدم التطابق 1% .
- مجموع التطابقات الأخرى 134 .
- نسبة التطابقات الأخرى 77 % .

خوارزميـة (Name Similarity) تـستخدم Dice Cofficient كخوارزميـة وزن، وخوارزميــة (LCS) Longest Common Subsequence ، و Levenshtein Distance . **والأسماء المدخلة للخوارزمية بطريقة (bigram model)**

التوصيات

1- العمل بخوارزميات أخرى كخوارزمية ***Porter Stemming Algorithm*** .

2- تطبيق هذه الخوارزميات على قواعد بيانات أخرى كقاعدة بيانات "**المكتبـات**" أو قاعـدة البيانات الخاصة بـ "**الضرائب**" وقد تم تطبيقها في ولاية تكساس الأمريكية وكان له الأثر الإيجابي بحيث أدت إلى ما قيمته على الأقل **(43)** مليون دولار زيادة في جمع الضرائب خلال **(18)** شهر.

3- استخدام مطابقة السلسة التقريبي في أهداف أخرى كـ "**تعـدين البيانـات** "، "**البحـث الوراثي**" (Genetic)، "**الاتصالات**".

المصـــادر

أ- المصادر العربية

1- خليل خليل، مبادئ قواعد البيانات العلائقية وطـرق تـصميمها، الموسـوعة العربيـة للكمبيوتر والانترنيت، (2005).

2- محمد، هيثم، فن المبيعات، منتديات الوعد الحق – القسم العلمي والتعليمي (2004).

ب - المصادر الأجنبية

3- Brown,G.M., The Approximate string Matching Engin,(2004).

4- Carstensen, A. ,"An Introduction to Double Metaphone and the principles behinds Soundex " , DMG ,(2007), Pp:2

5- D. Holmes and C. M. McCabe. Improving Precision and Recall For Soundex Retrieval.In Proceedings of the IEEE International conference on information technology-coding and computing (ITCC), Las Vages,(2002) .

6- Jizba,Richard, Measuring Search Effectiveness,(2000).

7- Michael Gilleland & Merriam Park, **Levenshtein Distance, in_ Three Flavors**,(2001),Pp2 .

8- Singh, B. S., **"Search Algorithms"**,DRTC, indian statistical institute Bangalore -560 059 ,(2003), Pp4-6 .
email:- brijesh@drtc.isibang.ac.in

9- Tindell Ralph, **Binary Search**,(2003),Pp1-13

10- White, T**., can't beat Jazzy**,(2004),Pp3 .
email:- tom@tiling.org.

ج- مواقع الانترنيت

11. www.cybrarians.info/news/salem.htm

ملحق (1)

جدول الـ Supplier		
اسم الحقل	النوع	الحجم
Supplier- ID	نص	10
Company- Name	نص	50
Contact- Person	نص	50
Contact- Title	نص	50
Address	نص	50
Email	نص	50
Phone	نص	20
Fax	نص	50
جدول الـ Products		
Product-Code	نص	10
Product-Name	نص	50
Unit-Price	عملة	
Units-In-Stock	رقم	عدد صحيح طويل
Recorder-Level	رقم	عدد صحيح طويل
جدول الـ Customers		
Customer- Code	نص	10
First - Name	نص	50
Last - Name	نص	50
Address	نص	50
Email	نص	50
Phone	نص	20
جدول الـ Order		
Order- No	نص	10
Customer- Code	نص	10
Date	تاريخ / وقت	
جدول الـ OrderDetails		
Order-No	نص	10
Product-Code	نص	10
Unit-Price	عملة	
Quantity	رقم	عدد صحيح طويل
Total-Price	عملة	
جدول الـ Deliver		
Deliver-No	نص	10
Supplier- ID	نص	10
Date	تاريخ / وقت	
جدول الـ DeliverDetails		
Deliver-No	نص	10
Product-Code	نص	10
Unit-Price	عملة	
Quantity	رقم	عدد صحيح طويل
Total-Price	عملة	