Matcher - מנוע זיהוי ישויות

מיקי קולקו

רקע

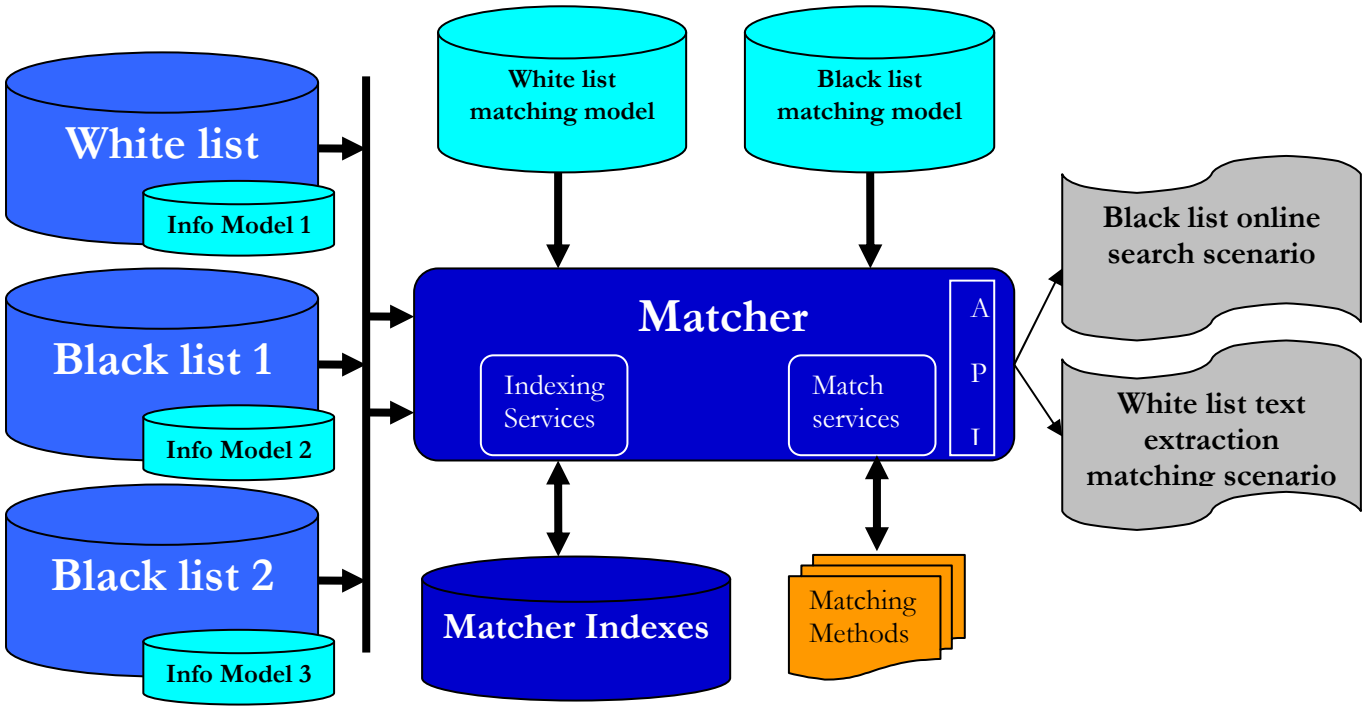
ה Matcher הינו מנוע זיהוי ישויות גמיש ורב עוצמה למצבים בהם המידע אינו מדויק ומתואם עם בסיסי המידע בארגון. המנוע משלב טכנולוגיית FUZZY MATCHING וארכיטקטורת אינדקסים ייחודית המאפשרת לארגונים לבנות יישומי זיהוי ואיתור ישויות בתחומי ביטחון, מודיעין, שירות לקוחות ועוד. המנוע מסייע בזיהוי והתאמת ישויות - אנשים וארגונים, תוך התגברות על חוסר דיוק ואי וודאות במידע.

ה Matcher נותן מענה למגוון רחב של תסריטי זיהוי של אנשים וארגונים כמו: חיפוש ישויות מול רשימות לבנות/שחורות, בדיקת תקינות של נתונים אשר חולצו ממידע לא מובנה כמו טקסט, סטנדרטיזציה של נתוני קלט או נתונים הנקלטים ממקורות חיצוניים וחיפוש מקוון מבוסס דמיון.

עקרונות הפתרון

ה Matcher הינו פתרון כללי למציאת דמיון בין ישויות בבסיסי נתונים ארגוניים. תהליך ההתאמה מתבצע באמצעות אלגוריתמים יעילים אשר מזהים אי התאמות במידע כמו טעויות איות, מידע חסר, חוסר סטנדרטיזציה הנובע מריבוי מקורות קלט , מידע המחולץ מטקסט, ריבוי סמנטיקות עסקיות ושילובים שונים של אי התאמות אלו.

הדיאגרמה הבאה מציגה שילוב של ה Matcher במערכת בקרת גבולות לזיהוי אנשים במגוון תסריטים:



ה Matcher תומך ביישומים מגוונים המנהלים מודל נתונים מגוון ומורכב. היישום ממפה את ישויות המידע באמצעות כלי מיפוי המותאם ל XML – מיפוי זה נקרא **מודל המידע** . מערכת האינדוקס ב Matcher עוקבת אחר השינויים במידע וע"פ **מודלי המידע** מסנכרנת את מערכת האינדקסים עם נתוני הארגון. תהליך הזיהוי וההתאמה

2001 Computers & System Services Ltd

22a' Raoul Wallenberg, Tel-Aviv, 69719 Israel ♣ Tel: 972-3-6474145 ♣Fax: 972-3-6474127 ♣ Email: officeh@int2001.co.il

מבוצע באמצעות מערך האינדקסים ביעילות ללא גישה לבסיס הנתונים. לוגיקת הזיהוי ניתנת להגדרה והתאמה מלאה של הלקוח באמצעות **מודל הזיהוי**. ניתן ליישם מספר מודלי זיהוי למגוון תסריטים אפשריים.

לדוגמא, מערכת בקרת גבולות בודקת נוסעים באמצעות חיפוש מקוון מבוסס דמיון ברשימה שחורה של חשודים:

	נתוני דרכון	**Black list**	**אלגוריתם לזיהוי**
שם	Muhamad Usman Abdel Raqeeb	Haj Mohd Othman Abdul Rajeeb	Multi culture name analysis
שם האב	Hasim	Chasim	Arabic phonetics similarity
גיל	19	20	Age range
אזרחות	Lebanon	Iran	Countries relationships

פרוט יכולות וערכי מוסף

מודל מידע גמיש התומך בעדכון בזמן אמת : ה Matcher מותאם לזהות ישויות ע"פ מודל מידע XMLי. באמצעות מודלי המידע וכלי המיפוי, מפתחי היישום ממפים את הישויות הרצויות למערך האינדקסים של ה Matcher. לאחר ביצוע המיפוי, ה Matcher עוקב אחר השינויים בזמן אמת באמצעות מנגנון תור בבסיס הנתונים. תור זה מאכלס את כל השינויים הנדרשים לסנכרון מערך האינדקסים.

מודל המידע המשמש את מערך הכנת האינדקסים זהה למודל המידע המשמש את תהליך הזיהוי. מודל זה מפשט את תהליך השילוב של ה Matcher והיישום באמצעות סכמה אחידה.

מודל זיהוי גמיש להתאמות היישום: תהליך הזיהוי נשלט במלואו ע"י היישום. מודל המידע מגדיר את שיטות הזיהוי לכל שדה ושדה, משקולות חשיבות בין השדות וספי זיהוי. פרמטרים אלו משמשים את ה Matcher לתכנן את אסטרטגיית הזיהוי האופטימלית לנתוני הקלט מחד ונתוני בסיס התונים מאידך. ניתן ליישם מספר מודלי זיהוי למגוון תסריטים אפשריים.

מערך מוכן של אלגוריתמים לזיהוי: ה Matcher משווק Out-of-the-box עם מערך אלגוריתמים לזיהוי סוגי נתונים שונים כמו: שמות, כתובות, תאריכים, מספרים מזהים. כל אלגוריתם לזיהוי מתמודד עם התאמות בהתאם לסוג הנתון. למשל התאמה של שם תכלול התאמה פונטית והתאמה של מספר דרכון יכלול החלפה של ספרות. כל אלגוריתם לזיהוי פועל באמצעות אינדקס מתאים המוגדר ומנוהל במערכת האינדקסים. למשל התאמה פונטית פועלת באמצעות אינדקס SOUNDEX. מערך האלגוריתמים ניתן להרחבה והתאמה לסוגי הנתונים הקיימים וסוגי נתונים חדשים.

ריבוי שפות למערכות גלובליות: ה Matcher תומך בזיהוי ישויות במאגרים מרובי שפות. המידע הארגוני מעובד ע"י מודול שפה ייעודי. מודול השפה כולל אלגוריתמים המתאמים לשפה ומאפשרים זיהוי מידע ע"פ אי התאמות המקובלות לשפה. מודוליי שפה נרכשים בנפרד ע"פ בחירת הלקוח.

זיהוי שמות מונחה תרבות: ה Matcher משולב עם הפתרון המוביל של חברת IBM לניתוח וזיהוי שמות – **IBM Globale Name Scorting**. טכנולוגיית הזיהוי של IBM משלבת מידע איכותי על כל שם ושם בהתאם לתרבות המוצא של השם. המידע על השם , הכולל ניתוח אוטומטי של תרבות השם, מין, חלוקת השם למרכיביו - מאפשר זיהוי ייחודי של אנשים וארגונים ברחבי העולם, ובמגוון תעתיקים. למשל השמות **Zhang Qiusu, Chang Ch'iu-Su, Chiusu Sae Chang, Cheung Yau So, Cheung Yau So** מתייחסים כולם לאותו שם שמקורו במזרח אסיה. מודול זה נרכש בנפרד ע"פ דרישת הלקוח.

זיהוי פונטי מרובה שפות: כחלק ממודול השפה מאפשר ה Matcher זיהוי פונטי רגיש שפה למידע טקסטואלי. הזיהוי הפונטי מאפשר איתור ישויות תוך התעלמות מאי התאמות מבוססות שמיעה וכתיב. אלגוריתם הזיהוי הפונטי הנו מבוסס חוקים המאפשרים גמישות והתאמות. ה Matcher כולל מענה לשפות אנגלית, עברית וערבית . שפות נוספות יתווספו ע"פ דרישה. למשל שם החברה האיראנית **Cobel Daron** נשמע כמו **Kobbel Daaron** בתעתיק לטיני.

זיהוי ישויות מספריות: בעולם הדיגיטלי נוצרים מקרים רבים בהם יש עיוותים בשדות מזהים כמו מספר דרכון, מספר רכב . ה Matcher כולל אלגוריתמים ייעודיים לאיתור דמיון בין מזהים הכוללים בין השאר החלפת ספרות, השמטת ספרות.

ביצועים מעולים מבוססי ארכיטקטורת אינדקסים ייחודית: מערך האינדקסים הנו לב ליבו של טכנולוגיית ה Fuzzy matching של ה Matcher. מערך האינדקסים ובשילוב רכיבי מטמון מאפשר זיהוי ישויות במאגרים רחבי היקף ובביצועים מעולים. מערך האינדקסים מנוהל ע"י ה Matcher באמצעות כלי ניהול המבצעים סנכרון עם שינויים בנתוני היישום .

Federation: למאגרים רחבי היקף, ולצורכי Scalability כולל ה Matcher מודול נפרד ל Federated Match. מודול זה מאפשר ביזור בסיס הנתונים למספר מערכי אינדקסים המפוזרים ע"פ פני מספר שרתים. הזיהוי המבוזר מתבצע במקביל וביעילות. מודול זה נרכש