Intihal algılama

İntihal tespiti, bir iş veya belge içerisinde intihal örneklerini bulma sürecidir. Bilgisayarların yaygın kullanımı ve İnternetin ortaya çıkışı, başkalarının çalışmalarını intihali kolaylaştırdı. Çoğu intihal vakası, belgelerin tipik olarak deneme veya rapor olduğu akademide bulunur. Bununla birlikte, intihal, romanlar, bilimsel makaleler, sanat tasarımları ve kaynak kodu dahil olmak üzere hemen her alanda bulunabilir. İntihal tespiti manuel veya yazılım destekli olabilir. El ile algılama, çok fazla belgenin ve mükemmel belleğin kullanılmasını gerektirir ve çok fazla belgenin karşılaştırılması gereken durumlarda ya da karşılaştırma için orijinal belgeler kullanılamıyorsa pratik değildir. Yazılım destekli algılama, çok sayıda belge koleksiyonunun birbiriyle karşılaştırılmasına olanak tanır ve bu da başarılı bir şekilde tespit edilmesini daha olası hale getirir.
Algılama yazılımından kurtulmak için yeterli kelime ikameleri kullanılarak intihal etme pratiği, rogeting olarak bilinir. [1]

Yazılım destekli algılama

Bilgisayar destekli intihal algılama (CaPD), intihal tespit sistemleri (PDS) olarak adlandırılan, özel IR sistemleri tarafından desteklenen bir Bilgi erişim (IR) görevidir.

Metin belgelerinde

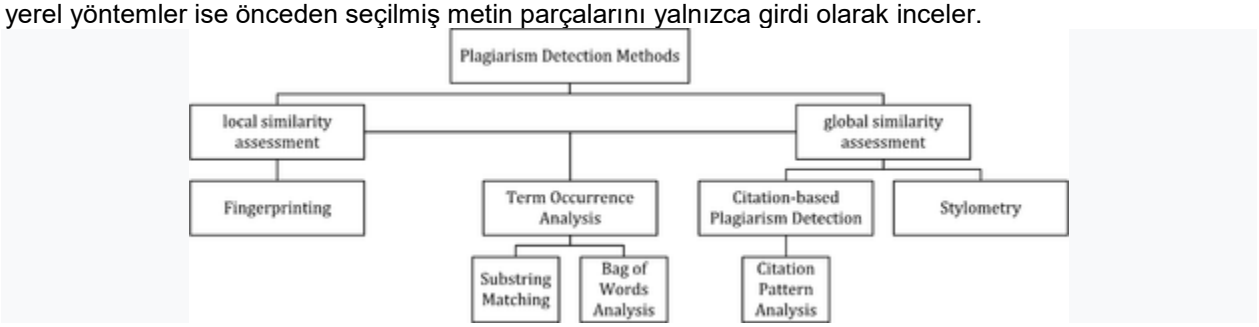
Metin-intihal tespitine yönelik sistemler, biri harici, diğeri de içsel olan iki genel algılama yaklaşımından birini uygular. [2] Harici algılama sistemleri, şüpheli bir belgeyi, orijinal olduğu varsayılan bir dizi belge olan bir referans koleksiyonuyla karşılaştırır. [3] Seçilen bir belge modeline ve önceden tanımlanmış benzerlik kriterlerine dayanarak, algılama görevi, şüpheli belgede metne seçilen bir eşiğin üzerindeki bir dereceye benzer metin içeren tüm belgeleri almaktır. [4] İçsel PDS, yalnızca harici belgelere karşılaştırma yapmadan değerlendirilecek metni analiz eder. Bu yaklaşım, bir yazarın eşsiz yazma stilindeki değişiklikleri potansiyel intihal göstergesi olarak tanımlamayı amaçlamaktadır. [5] PDS, insan yargısı olmaksızın intihali güvenilir bir şekilde belirleyememektedir. Benzerlikler önceden tanımlanmış belge modelleri yardımıyla hesaplanır ve yanlış pozitifleri temsil edebilir. [6] [7] [8] [9] [10]

Yükseköğretim ortamlarında etkinlik

Bir yüksek öğretim ortamında intihal tespit yazılımının etkinliğini test etmek için bir çalışma yapılmıştır. Çalışmanın bir bölümü bir grup öğrencinin bir kâğıt yazmasını istedi. Bu öğrenciler ilk olarak intihal konusunda eğitildiler ve çalışmalarının bir intihal tespit sistemi ile yürütüleceğini öğrendiler. İkinci bir grup öğrenci intihal hakkında bilgi sahibi olmayan bir makale yazmak için görevlendirildi. Araştırmacılar grupta daha düşük oranlar bulmayı beklerken, her iki grupta da kabaca aynı miktarda intihal bulmuşlardır. [11]

Yaklaşımlar

Aşağıdaki şekil, bilgisayar destekli intihal tespitinde kullanılmakta olan tüm algılama yaklaşımlarının bir sınıflandırmasını temsil etmektedir. Yaklaşımlar, üstlendikleri benzerlik değerlendirmesinin türü ile karakterize edilir: küresel veya yerel. Küresel benzerlik değerlendirme yaklaşımları, metnin daha büyük bölümlerinden veya bir bütün olarak belgenin benzerliğini hesaplamak için kullanılan özellikleri kullanır, yerel yöntemler ise önceden seçilmiş metin parçalarını yalnızca girdi olarak inceler.



Bilgisayar destekli intihal tespit yöntemlerinin sınıflandırılması

Parmak İzi

Parmak izi şu anda intihal tespitine en çok uygulanan yaklaşımdır. Bu yöntem, birden çok alt dizgi (n-gram) grubunu seçerek, belgelerin özet özetlerini oluşturur. Setler parmak izlerini temsil eder ve elementlerine minutia denir. [12] [13] Şüpheli bir belge, parmak izi hesaplanarak ve referans koleksiyonunun tüm dokümanları için önceden hesaplanmış bir parmak izi dizisi ile minutia sorgulama yoluyla intihal için kontrol edilir. Diğer belgelerdekilerle eşleşen Minutiae, paylaşılan metin parçalarını gösterir ve seçilmiş bir benzerlik eşiğini aşarsa potansiyel intihal önerir. [14] Hesaplama kaynakları ve zaman, parmak izi ile ilgili faktörleri sınırlandırmaktadır; bu nedenle, bu yöntem tipik olarak yalnızca minutia'nın bir alt kümesini hesaplamayı hızlandırmak ve Internet gibi çok büyük koleksiyonlarda çeklere izin vermek için karşılaştırır. [12]

Dize eşleme

Dize eşleştirmesi bilgisayar bilimlerinde kullanılan yaygın bir yaklaşımdır. İntihal tespiti problemine uygulandığında, sözel metin çakışmaları için belgeler karşılaştırılır. Bazıları dış intihal tespitine adapte edilmiş olan bu görevi ele almak için çok sayıda yöntem önerilmiştir. Şüpheli bir belgenin bu ayarda kontrol edilmesi, referans koleksiyonundaki tüm belgeler için, karşılaştırılabilir bir şekilde karşılaştırılabilir temsillerin hesaplanmasını ve saklanmasını gerektirir. Genel olarak, sonek ağaçları veya sonek vektörleri gibi belge modelleri, bu görev için kullanılmıştır. Bununla birlikte, alt dizgi eşleşmesi hesaplama pahasına pahalıdır, bu da onu büyük doküman koleksiyonlarını kontrol etmek için uygun olmayan bir çözüm haline getirir. [15] [16] [17]

Kelime çantası

Kelime analizi çantası, intihal tespitinin alanına, geleneksel bir IR kavramı olan vektör uzayının kabul edilmesini temsil eder. Dokümanlar, bir veya birden fazla vektör olarak temsil edilir, ör. Çift bilge benzerlik hesaplamaları için kullanılan farklı belge parçaları için. Benzerlik hesaplaması, daha sonra geleneksel kosinüs benzerlik ölçüsüne veya daha karmaşık benzerlik ölçütlerine dayanabilir. [18] [19] [20]

Atıf analizi

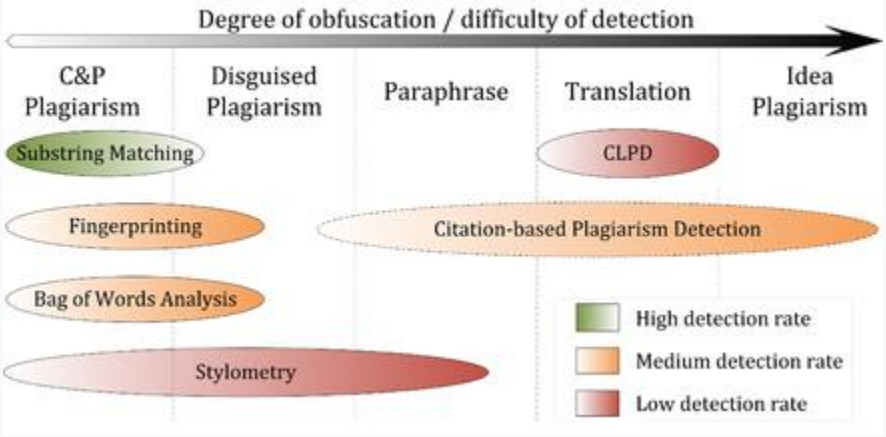
Atıfta dayalı intihal tespit (CbPD) [21] alıntı analizine dayanır ve metinsel benzerliğe dayanmayan intihal tespitinin tek yaklaşımıdır. [22] CbPD, alıntı sekanslarındaki benzer modelleri tanımlamak için alıntı ve referans bilgisini metinlerde inceler. Bu yaklaşım, bilimsel metinler veya alıntı içeren diğer akademik belgeler için uygundur. İntihal tespit etmek için alıntı analizi nispeten genç bir kavramdır. Ticari yazılım tarafından kabul edilmemiştir, ancak atıf temelli bir intihal tespit sisteminin ilk prototipidir. [23] İncelenen belgelerde benzer sıralama ve alıntıların yakınlığı, alıntı deseni benzerliklerini hesaplamak için kullanılan temel kriterlerdir. Atıf paternleri, karşılaştırılan belgelerle paylaşılan alıntıları münhasıran içeren alt dizileri temsil eder. [22] [24] Modelde paylaşılan alıntıların mutlak sayısı veya nispi fraksiyonu gibi faktörlerin yanı sıra, bir belgede birlikte atıfta bulunma olasılığının da, modellerin benzerlik derecesini nicelleştirdiği düşünülmektedir. [22] [24] [25] [26]

Stilometri

Stilometri, bir yazarın benzersiz yazma stilini [27] [28] ölçmek için istatistiksel yöntemler kullanır ve esas olarak yazarlık ilişkilendirmesi veya içsel CaPD için kullanılır. Farklı metin bölümleri için stilometrik modellerin oluşturulması ve karşılaştırılmasıyla, diğerlerinden stilistik olarak farklı olan ve potansiyel olarak intihal olan geçişler tespit edilebilir. [5]

Performans

İntihal tespit sistemlerinin karşılaştırmalı değerlendirmeleri [3] [29] [30] [31] [32] [33], performanslarının mevcut intihal türüne bağlı olduğunu göstermektedir (şekle bakınız). Alıntı örüntü analizi haricinde, tüm algılama yaklaşımları metinsel benzerliğe dayanır. Bu nedenle, saptama doğruluğunun azalması, daha fazla intihal vakalarının gizlendiğini belirtmektedir.



Var olan intihal tipine bağlı olarak CaPD yaklaşımlarının tespit performansı.

Aslına uygun kopyalar, aka kopyalama ve yapıştırma (c & p) intihali veya başarılı bir şekilde gizlenmiş intihal vakaları, eğer kaynak yazılım tarafından erişilebilir ise mevcut harici PDS tarafından yüksek doğrulukla tespit edilebilir. Özellikle temel eşleştirme prosedürleri, c & p intihalciliği için iyi bir performans sağlar, çünkü bunlar genellikle son ek ağaçları gibi kayıpsız belge modelleri kullanır. Kopyaların tespitinde parmak izi ya da kelime çantası modelinin kullanıldığı sistemlerin performansı, kullanılan belge modelinin maruz kaldığı bilgi kaybına bağlıdır. Değişebilen parcalama ve seçme stratejileri uygulayarak, temel eşleştirme prosedürleriyle karşılaştırıldığında, gizli intihal halindeki vasat formları tespit etmede daha iyilerdir.

Stilometriyi kullanan asıl intihal saptaması, dilsel benzerliği karşılaştırarak bir dereceye kadar metinsel benzerlik sınırlarının üstesinden gelebilir. İntihal edilmiş ve orijinal bölümler arasındaki biçimsel farklılıkların anlamlı olduğu ve güvenilir bir şekilde tanımlanabildiği göz önüne alındığında, stilometri gizlenmiş ve başka kelimelerle yorumlanan intihal tespitinde yardımcı olabilir. Segmentlerin, fikir hırsızlığı yapan kişinin kişisel yazma tarzına daha çok benzediği noktaya kuvvetle ifade edildiği durumlarda veya bir metin çok sayıda yazar tarafından derlenmiş olduğu durumda, stilometrik karşılaştırmalar başarısız olabilir. Stein tarafından yapılan deneyler [34] gibi, 2009, 2010 ve 2011 yıllarında düzenlenen İntihal Saptama hakkındaki uluslararası yarışmaların sonuçları [3] [32] [33] stilometrik analizin sadece birkaç bin uzunluğundaki belgeler veya on binlerce kelime için güvenilir bir şekilde çalıştığını yani yöntemin uygulanabilirliğini CaPD ayarlarına sınırladığını göstermektedir. Çevrilmiş intihalleri tespit edebilen yöntemler ve sistemler üzerinde artan miktarda araştırma yapılmaktadır. Halihazırda, çapraz-dil intihal tespiti (CLPD) olgun bir teknoloji olarak görülmemektedir [35] ve ilgili sistemler pratikte tatmin edici tespit sonuçları elde edememiştir. [31]

Atıf biçim analizi kullanılarak yapılan alıntı temelli intihal saptaması, diğer algılama yaklaşımlarıyla kıyaslandığında daha güçlü izahları ve çevirileri daha yüksek başarı oranlarıyla tanımlayabilir, çünkü metinsel özelliklerden bağımsızdır.[22][25] Bununla birlikte, alıntı biçim analizi, yeterli alıntı bilgisinin varlığına bağlı olduğundan, akademik metinlerle sınırlıdır. Kopyala-yapıştır veya salla-yapıştır tarzındaki intihal durumlarında tipik olan izinsiz alıntı yapılan daha kısa bölümlerde metin tabanlı yaklaşımlardan daha düşüktür; sonuncusu, farklı kaynaklardan biraz değiştirilmiş parçaların karıştırılması anlamına gelir. [36].

Yazılım [edit]

Metin belgelerinde kullanılmak üzere intihal tespit yazılımının tasarımı bir dizi etkenle karakterize edilir: [kaynak belirtilmeli]

Etken	Tanım ve alternatifler

Araştırma kapsamı	Halka açık internette, arama motorlarını / Kurumsal veritabanlarını / Yerel, sisteme özel veritabanını kullanma. [kaynak belirtilmeli]
Analiz süresi	Bir dokümanın gönderildiği zaman ile sonuçların elde edilme zamanı arasındaki gecikme. [kaynak belirtilmeli]
Belge kapasitesi / Toplu işlem	Sistemin zaman birimi başına işleyebileceği belge sayısı. [kaynak belirtilmeli]
Yoğunluk kontrolü	Sistem, ne sıklıkta ve hangi tür doküman parçaları için (paragraflar, cümleler, sabit uzunluklu kelime dizileri) arama motorları gibi harici kaynakları sorgular.
Karşılaştırma algoritması türü	Sistemin belgeleri birbiriyle karşılaştırmak için kullandığı yolu tanımlayan algoritmalar. [kaynak belirtilmeli]
Hassasiyet ve Hatırlama	İşaretli belgelerin toplam sayısına oranla intihal olarak doğru bir şekilde işaretlenen doküman sayısı ve aslında intihal edilen belgelerin toplam sayısı. Yüksek hassasiyet, birkaç yanlış pozitifin bulunduğu ve yüksek anımsatmanın birkaç yanlış negatifin fark edilmeden kaldığı anlamına gelir. [kaynak belirtilmeli]

Çoğu büyük ölçekli intihal tespit sistemleri, analiz için sunulan her ek belge ile birlikte büyüyen büyük, dahili veritabanlarını (diğer kaynaklara ek olarak) kullanır. Bununl yanı sıra, bu özellik, bazıları tarafından öğrenci telif hakkının ihlali olarak kabul edilir. [kaynak belirtilmeli]

Kaynak kodda [Düzenle]

Bilgisayar kaynak kodundaki intihal da sık görülür ve belgede metin karşılaştırmaları için kullanılanlardan farklı araçlar gerektirir. Akademik kaynak kodlu intihal konusunda önemli araştırmalar yapılmıştır. [37]

Kaynak-kod intihalinin ayırt edici bir yönü, geleneksel intihalde bulunabilecek bir makale fabrikası olmamasıdır. Çoğu programlama ödevleri, öğrencilerin çok özel gereksinimlere sahip programlar yazmasını beklediğinden, halihazırda bunları karşılayan mevcut programları bulmak çok zordur. Harici kodun bütünleştirilmesi genellikle sıfırdan yazmaktan daha zor olduğundan, çoğu intihalci öğrenciler bunu akranlarından seçer.

Roy ve Cordy'ye göre, [38] kaynak kod benzerliği algılama algoritmaları şunlara göre sınıflandırılabilir:

- Dizeler - bölümlerin tam metinsel eşleşmelerini aramak, örneğin beş kelimelik çalışmalar. Hızlı, ancak tanımlayıcıları yeniden adlandırmada karışabilir.
- Token kuponları – dizelerle olduğu gibi, ancak önce programı token kuponlarına dönüştürmek için bir lexer kullanarak. Bu, beyaz boşlukları, yorumları ve tanımlayıcı adlarını ayırır ve basit metin değiştirmeleri için sistemi daha sağlam hale getirir. Çoğu akademik intihal tespit sistemi, token dizileri arasındaki benzerliği ölçmek için farklı algoritmalar kullanarak bu düzeyde çalışır.

- • Parçalama Ağacı - parçalama ağaçlarını oluşturma ve karşılaştırma. Bu, daha yüksek düzeydeki benzerliklerin tespit edilmesini sağlar. Örneğin, ağaç karşılaştırması koşullu ifadeleri normalleştirebilir ve birbirine benzer yapıları eşdeğer olarak tespit edebilir.
- • Program Bağımlılık Grafikleri (PDGs) – Bir PDG, bir programdaki gerçek kontrol akışını yakalar daha büyük bir zorluk ve hesaplama süresi pahasına, daha yüksek seviyeli denklemlerin yerleştirilmesine izin verir.
- • Metrikler - metrikler, kod bölümlerinin belirli kriterlere göre 'puanlarını' yakalar; örneğin, "döngüler ve koşulların sayısı" veya "kullanılan farklı değişkenlerin sayısı". Metriklerin hesaplanması basittir ve hızlı bir şekilde karşılaştırılabilir, ancak yanlış pozitif sonuçlara da yol açabilir: bir grup metrikte aynı puanlara sahip iki fragman tamamen farklı şeyler yapabilir.
- • Hibrid yaklaşımlar - örneğin, parçalama ağaçları + son ek ağacı, bir dizi eşleşmeli veri yapısı olan parçalama ağaçlarının algılama yeteneğini, son ek ağaçlarının oluşturduğu hızla, birleştirebilir.

Bir önceki sınıflandırma, kodun yeniden düzenlenmesi için geliştirilmiştir ve akademik intihal tespitine yönelik değildir (yeniden düzenlenmenin önemli bir amacı, literatürde kod klonları olarak anılan çift koddan kaçınmaktır). Yukarıdaki yaklaşımlar benzerliğin farklı seviyelerine karşı etkilidir; yüksek seviyeli benzerlik benzer tanımlamalara bağlı olabilirken düşük seviyeli benzerlik, aynı metne karşılık gelir. Akademik bir ortamda, tüm öğrencilerin aynı tanımlamalara göre kodlaması beklendiğinde, işlevsel olarak eşdeğer kod (üst düzey benzerlik ile) tamamen beklenir ve yalnızca düşük düzeydeki benzerlik kopya kanıtı olarak kabul edilir.